**Proceedings Paper:**
Park, C. and Hain, T. orcid.org/0000-0003-0939-3464 (2025) Semi-supervised learning for automatic speech recognition with word error rate estimation and targeted domain data selection. In: Scharenborg, O., Oertel, C. and Truong, K., (eds.) Proceedings of Interspeech 2025. Interspeech 2025, 17-21 Aug 2025, Rotterdam, The Netherlands. International Speech Communication Association (ISCA) , pp. 3663-3667. ISSN: 2958-1796 EISSN: 2958-1796

https://doi.org/10.21437/Interspeech.2025-191

# Semi-Supervised Learning for Automatic Speech Recognition with Word Error Rate Estimation and Targeted Domain Data Selection

*Chanho Park, Thomas Hain*

School of Computer Science, University of Sheffield, UKUniversity of Sheffield, UK

cpark12@sheffield.ac.uk, t.hain@sheffield.ac.uk

## Abstract

There is a growing demand for leveraging untranscribed multi-domain data in semi-supervised learning (SSL) for automatic speech recognition (ASR) to broaden its applications. However, domain mismatch between source and target data can limit SSL's performance gains, even when transcript accuracy for training is high. While word error rate (WER) estimation (WE) methods for automatic transcription have advanced, they remain insufficient for handling multi-domain data.

This paper proposes a novel data selection method for SSL in ASR that integrates WE and acoustic domain similarity (ADS). For WE, multi-target regression for error rate prediction (MTR-ER) is introduced, while ADS is incorporated as a selection criterion, measured using noise-contrastive estimation. The effectiveness of this approach is demonstrated through comparisons with a confidence-based method. Results show that combining WE and ADS achieves 26.66% of the expected performance improvement of fully supervised learning.

**Index Terms**: speech recognition, semi-supervised learning, word error rate estimation, acoustic domain similarity

## 1. Introduction

As data-driven approaches to automatic speech recognition (ASR) have become widely adopted for ASR, the demand for manual transcripts for training has increased to more than a thousand hours [1, 2]. However, manual transcription still remains costly and time-consuming and it is not available for all domains, e.g., different recording conditions or topics. As an alternative to the approaches that rely on large amounts of manually transcribed data, researchers have explored semi-supervised learning (SSL), leveraging untranscribed utterances to enhance ASR performance [3, 4, 5, 6, 7]. One of the early attempts at SSL was proposed in [8]. The authors used an ASR system, known as a seed model, to transcribe utterances and generate transcripts, referred to as ASR transcripts. They found that ASR transcripts with up to 20% error could still improve the performance of an ASR system when the transcripts were used for training. To ensure the high accuracy of these ASR transcripts, training data were selected using information from ASR decoding, such as a confidence score.

Although this method has proven to be useful for data selection, it was reported that overconfident data contribute minimally to ASR performance improvement [9, 10]. Moreover, methods relying on confidence scores not only require access to the internal process of the ASR system, which is not always possible, especially in industrial systems, but also overlook deletions in ASR system output. To address these issues, a word error rate (WER) estimation (WE) method for an ASR system's output using features obtainable without ASR decoding was proposed in [11, 12]. Recently, WER estimation models using self-supervised learning representations (SSLRs) for speech and text was introduced [13, 14]. Fe-WER [14], in particular, adopted the temporal mean of speech representations over time to improve computational efficiency without sacrificing WE performance. Building on these advancements, the Fe-WER model was employed to select ASR transcriptions estimated to have high accuracy for training. However, the performance was found to degrade on out-of-domain data, as will be shown in this paper. To alleviate this issue, this paper proposes a method that trains a WE model to predict more detailed error information, including substitution, deletion and insertion error rates (ERs) alongside WER. This approach leverages multiple real-valued outputs from the ER estimation model to improve generalisation by simultaneously learning multiple related tasks, a method known as multi-target regression (MTR) [15, 16].

Nevertheless, when training data are selected from heterogeneous domains, high accuracy of ASR transcription is not sufficient. Domain mismatch between source and target domains can lead to performance degradation, especially between read speech and telephone speech [17]. Moreover, findings from [18] suggest that domain mismatch can have a greater impact on ASR performance than the quantity of training data. To address this, acoustic domain similarity (ADS) is introduced as an additional criterion for data selection, building upon the unsupervised data selection method in [19], which accounts for matching domains. This similarity is measured using mutual information modelled through an objective function for noise-contrastive estimation (NCE) [20] with target data.

In this paper, a data selection method for SSL in ASR is proposed that incorporates WE and ADS. First, a WE model with MTR for ER prediction (MTR-ER) is used to select highly accurate ASR transcripts. Second, ADS between an utterance and target data is measured using a loss based on NCE. By incorporating these techniques, data can be effectively selected from an untranscribed multi-domain data pool for SSL in ASR.

The contributions of this work are as follows:

- Proposal of a new data selection method for SSL in ASR, utilising an untranscribed multi-domain data pool
- Extensive experimental evaluation, including analysis of different amounts of selected data, comparison with a confidence-based method and random selection
- Demonstration of effectiveness, showing that combining these two criteria yields 26.66% of the performance gain expected from fully supervised learning

## 2. Semi-Supervised Learning with Word Error Rate Estimation and Acoustic Domain Similarity

This section outlines the proposed SSL approach, which leverages WER estimation and acoustic domain similarity (ADS) to enhance ASR performance in multi-domain settings. The overall steps of SSL for ASR with WE and targeted domain data selection are as follows:

1. An ASR model $M_{\text{ASR}}$, a WE model $M_{\text{WE}}$, a ADS model $M_{\text{ADS}}$, a dataset with reference transcripts $D_R$, an untranscribed data pool $D_U$, a target dataset $D_T$.
2. Train $M_{\text{ASR}}$ with $D_R$.
3. Generate $D_A$ by applying $M_{\text{ASR}}$ to $D_U$.
4. Build $M_{\text{WE}}$ with $D_R$ and determine a threshold $\tau_W$
5. Select $D_W$ from $D_A$ using $M_{\text{WE}}$ with $\tau_W$.
6. Build $M_{\text{ADS}}$ with $D_T$ and determine a threshold $\tau_S$
7. Select $D_S$ from $D_A$ using $M_{\text{ADS}}$ with $\tau_S$.
8. Build $M'_{\text{ASR}}$ with $D_R \cup \{D_W \cap D_S\}$.

### 2.1. Word Error Rate Estimation

This section describes the WER estimation process, which aims to predict various error rates of ASR transcripts. WER is defined as the ratio of substitution $S$, deletion $D$ and insertion $I$ errors in an ASR transcript and a total number of words $N$ in a reference transcript: WER $= (S+D+I)/N$. The rate of each error type is also defined as its ratio to the total word count, $N$. For example, substitution error rate $y_S$ is defined as $\frac{S}{N}$. WER of ASR transcripts is estimated using a model with MTR for ER estimation, MTR-ER, as illustrated in Figure 1. This model

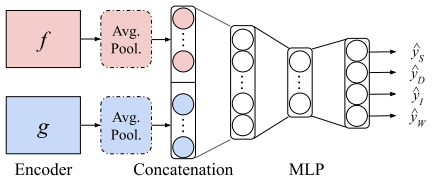



Figure 1: *Multi-target regression for error rate prediction*

consists of encoders, $f(\cdot)$ and $g(\cdot)$ for speech and text, respectively, along with a multi-layer perceptron (MLP) for WE. The estimation is given by:

$$\widehat{\mathbf{y}}^l = \text{MLP}(\text{concat}[\text{avg}(f(X^l)); \text{avg}(g(A^l))]) \quad (1)$$

where $X^l$ and $A^l$ represent the utterance and the ASR transcript of the $l$-th pair, respectively, and $\widehat{\mathbf{y}}$ is estimation of insertion $\widehat{y}_I$, deletion $\widehat{y}_D$, substitution $\widehat{y}_S$ and word error rates $\widehat{y}_W$ of $A$. avg$(\cdot)$ is an average pooling function. The model is trained by minimising the sum of mean squared error (MSE) across all error rates:

$$\text{MSE} = \frac{1}{M}\sum_{l=1}^{M}\sum_{j\in\{I,D,S,W\}}(y_j^l - \widehat{y}_j^l)^2 \quad (2)$$

where $M$ is the total number of training instances and $y$ is an actual error rate. Root mean square error (RMSE) and Pearson Correlation Coefficient (PCC) were used to evaluate both the accuracy and reliability of WER predictions. PCC measures the linear relationship between two variables, ranging from -1 (negative correlation) to 0 (no linear relationship) to 1 (positive correlation).

### 2.2. Acoustic Domain Similarity

To complement WER estimation, ADS is introduced to measure the similarity between utterances in a data pool $D_U$ and a target dataset $D_T$, ensuring that the selected data aligns closely with the target domain. ADS is calculated for each utterance using the mutual information between a context representation $\mathbf{c}_t = g(\mathbf{z}_t \dots \mathbf{z}_{t-v})$, where $t$ represents the current time and $v$ represents the context length, and a latent representation $\mathbf{x}_{t+k}$ of a future observation in speech, where $k$ is the future step. This mutual information can be modelled with $D_T$ using an objective function for noise-contrastive estimation [21].

$$\mathcal{L}_k(\mathbf{x}_t) = -[\log(f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)) + \sum_{\tilde{\mathbf{x}}\in p_n}\log(1 - f_k(\tilde{\mathbf{x}}, \mathbf{c}_t))] \quad (3)$$

where $f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)$ is a density ratio preserving the mutual information between $\mathbf{x}_{t+k}$ and $\mathbf{c}_t$, and $p_n$ is a set of random samples from the same recording. The utterance-level loss of a sequence $X_{1:T} = \{\mathbf{x}_1 \dots \mathbf{x}_t \dots \mathbf{x}_T\}$ is given as:

$$\mathcal{L}(X_{1:T}) = \frac{1}{T-k}\sum_{t=1}^{T-k}\sum_{k=1}^{K}\mathcal{L}_k(\mathbf{x}_t) \quad (4)$$

where $K$ is the maximum number of future steps. ADS is then computed as:

$$\text{ADS}(X^l) = \frac{\frac{1}{|D_T|}\sum_{X^m\in D_T}\mathcal{L}(X^m)}{\mathcal{L}(X^l)} \quad (5)$$

where $X^l$ is the $l$-th utterance in the data pool $D_U$ and $|D_T|$ is the number of utterances in the target data $D_T$. The average loss of target data is divided by the utterance-level loss for normalisation of the similarity value so that it increases as the mutual information increases and the utterance-level loss decreases. If the utterance loss is close to the average loss of the target data, the value approaches 1.

### 2.3. Data Selection for Semi-supervised Learning

WE and ADS were used to select training data for SSL. Thresholds for the measures of WE and ADS are denoted as $\tau_W$ and $\tau_S$, respectively. The pairs of utterances and transcripts whose estimate of WER is lower than $\tau_S$ were selected, while the utterances whose domain similarity was lower than the threshold were filtered out.

$$D_W = \{X^l, A^l\}, \text{where } \widehat{y}_W^l < \tau_W$$
$$D_S = \{X^l, A^l\}, \text{where ADS}(X^l) > \tau_S$$

The data selection methods are evaluated using WER Recovery [22], which measures the ratio of ASR performance gain achieved through SSL to supervised learning for ASR.

## 3. Experimental Setup

### 3.1. Datasets

CHiME-5 (CH5) [23] was used as a target domain dataset, consisting of 50 hours of conversational speech recorded in a home environment. Among the multiple channels, the binaural microphone recording of speakers was used. An untranscribed data pool was composed of spoken utterances from five ASR corpora. The AMI corpus [24] comprises meeting recordings involving up to four participants in an office environment. A subset of AMI, Full-corpus-ASR, was used in this experiment.

LibriSpech (LSP) [25] offers approximately 1000 hours of read speech from books. The Switchboard (SWB) corpus [26] encompasses two-sided telephone conversations. Ted-Lium 3 (TL3) [27] is a corpus of 452 hours of audio talks, and the Wall Street Journal (WSJ) corpus [28] is composed of read speech with machine-readable texts from Wall Street Journal news articles. The hours of the datasets are summarised in Table 1.

Table 1: *Hours of Speech in ASR corpora used.*

| | Corpus | Training | Dev | Test |
|---|---|---|---|---|
| Target | CH5 | 37.76 | 6.07 | 5.37 |
| Data Pool | AMI | 64.80 | 5.12 | 8.68 |
| | LSP | 961.25 | 10.51 | 10.75 |
| | SWB | 311.26 | 4.61 | 4.59 |
| | TL3 | 444.62 | 6.13 | 3.57 |
| | WSJ | 81.485 | 2.20 | 2.22 |

### 3.2. Automatic Speech Recognition Models

A HuBERT large model[1] pre-trained on 60k hours of Libri-Light [29] was fine-tuned according to the publicly released recipe. The seed model was fine-tuned on CH5 Training, while SSL models were fine-tuned on both the CH5 Training set and automatically transcribed data selected using the proposed method. All models were trained to the point of overfitting, with a limit of 3720 epochs. Default hyper-parameters were used, except for the masked input length, which was set to 2 to accommodate short utterances. All transcripts were standardised using the Whisper normaliser [2] and NeMo text processing tool [3].

### 3.3. Word Error Rate Estimation Models

For WE, the Fe-WER model was compared with MTR-ER, which estimates insertion, deletion, substitution ERs and WER. Both models were trained on the 32 hours of data in CH5 Training with their corresponding ASR transcripts generated using the seed model. Then they were evaluated on out-of-domain datasets. To minimise data imbalance, the number of ASR transcripts with a WER of 0 was limited to the sum of the second and third most frequent bins in a 100-bin histogram. Features for speech and text were extracted using HuBERT large and XLM-R large models. Hyper-parameters were determined via grid search, with input and output layers fixed at 2048 and 1, respectively. The sizes of two hidden layers were chosen from the ranges [300, 600, 900] and [16, 32, 64], respectively. An Adam optimiser with a Cosine Annealing scheduler[4] was used and learning rates were selected from [1e-4, 3e-4, 7e-4, 1e-3, 3e-3, 7e-3].

### 3.4. Acoustic Domain Similarity

For ADS computation, a wav2vec [20] model was employed, where the density ratio $f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)$ was implemented. Using the default hyper-parameters of the wav2vec large [5], kernel sizes were set to (16, 16) and strides to (8, 8). The context networks were comprised of three layers with increasing kernel sizes (2, 3, 4) to fit the smaller amount of data in $D_T$. Training stopped after 15 epochs of no performance improvement on a validation set. This model was pre-trained on CH5 Test; then the ADS of each utterance in AMI, LSP, SWB, TL3 and WSJ Training was computed.

[1] https://github.com/facebookresearch/fairseq
[2] https://github.com/openai/whisper
[3] https://github.com/NVIDIA/NeMo-text-processing
[4] https://pytorch.org/docs/stable/optim.html
[5] https://github.com/facebookresearch/fairseq

### 3.5. Data Selection for Semi-supervised Learning

$\tau_W$ was set at 20%, which was below the ASR model's performance with reference transcripts (see Figure 2). For ADS, $\tau_S$ of 0.74 was determined from the distribution of ADS in CH5 Test (see Figure 5), as it was assumed that some target data might also be outliers in terms of acoustic domain similarity. Therefore, data with acoustic domain similarity below the maximum similarity of the bottom 10% in the target dataset were removed. Additionally, to investigate performance improvement based on the amount of selected data, 8 and 32 hours of data were chosen. To control the quantity of selected data, utterances were sampled from $\{D_W \cap D_S\}$ in ascending order of WER estimates.

For comparison, data were also selected using a random strategy as well as a weighted confidence scoring method (WCS). WCS calculates the average of softmax probabilities produced by the seed model at the token level.

## 4. Results

### 4.1. A Seed Model

As a result of comparison of decoding strategies, Viterbi decoding was adopted for better performance across multiple domains on average. A HuBERT model trained on 32 hours of data was selected as the seed model to confirm that further improvement is possible with additional data as show in Figure 2.

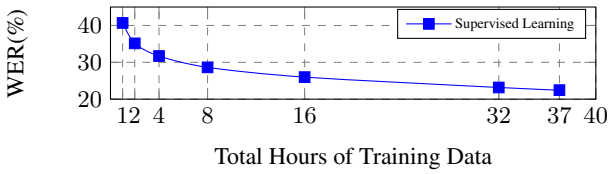



Figure 2: *Performance of HuBERT fine-tuned on CH5 Training*

### 4.2. Word Error Rate Estimation

To compare Fe-WER and MTR-ER's performance on out-of-domain datasets, results excluding CH5 were averaged. The MTR-ER model achieved a lower RMSE of 0.2153 compared to 0.2174 for the Fe-WER model, indicating a slight improvement in error minimization. Additionally, MTR-ER obtained a higher PCC of 0.6286, surpassing the Fe-WER model's PCC of 0.6044. These results suggest that MTR-ER offers better generalisation on out-of-domain data, with improved accuracy and correlation metrics.

Figure 3 shows that the cumulative mean of the WER reference continuously increases when ASR transcripts are selected in ascending order of MTR-ER's estimates. In contrast, the cumulative mean of the WER reference initially spikes and then decreases until reaching 8%, after which it gradually increases. The spike in the low range of WER estimate could be critically harmful on the ASR performance and is a weakness in selecting a small amount of data. Based on these results, MTR-ER was used to select low-WER transcripts for ASR training.

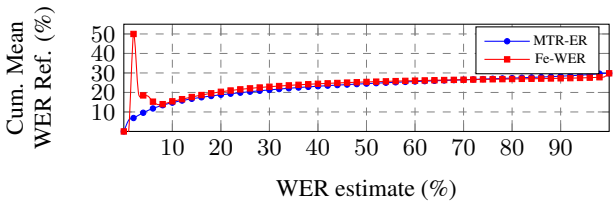
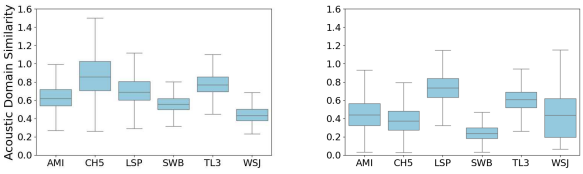



Figure 3: *Cumulative mean of reference WER according to a threshold for WER estimate.*

### 4.3. Acoustic Domain Similarity

The ADS distribution is plotted in Figure 4. The box plots show the distribution of ADS values in the data pool when the target data are CH5 Test and LSP Test. The mean ADS for CH5 is the highest in Figure 4a, while the mean ADS for LSP is the highest in Figure 4b. This observation indicates that the distributions are influenced by the target data. Based on the mean ADS, TL3 is the most similar to CH5, while WSJ is the most dissimilar. This result is as expected because WSJ consists of read speech recorded in studios, while CH5 comprises conversational speech recorded during dinner.




(a) *CH5 Test* (b) *LSP Test (clean)*

Figure 4: *ADS distributions when target data are CH5 Test and LSP Test (clean).*

The distribution of ADS in the data pool is relatively left-skewed as shown in Figure 5. The majority of utterances have ADS values below 1, which represents the mean ADS of the utterances in the target data. The dotted line indicates the maximum value of the 10th percentile, used as a threshold to remove dissimilar domain data.

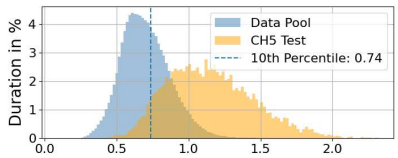



Figure 5: *Distribution of ADS in the data pool.*

### 4.4. Data selection

The statistics of the data selected using random, WCS and the proposed method are summarised in Table 2. The proposed method yields the lowest average WER reference on all the datasets. When data were selected using WCS, over twenty thousands SWB utterances were chosen, in an average WER reference of 76.21%. The proposed method selected fewer SWB utterances than the other methods due to their high WER. Additionally, it selected fewer WSJ utterances due to their low ADS values, despite having the lowest average WER estimates.

Table 2: *Statistics for 32 hours of selected data.*

| Data Selection | Corpus | # Utt. | Dur. (h) | Avg. WER Ref. | Avg. WER Est. | Avg. ADS |
|---|---|---|---|---|---|---|
| Random | AMI | 1557 | 1.13 | 36.23% | - | 0.63 |
| | LSP | 4831 | 16.59 | 13.10% | - | 0.72 |
| | SWB | 4190 | 5.24 | 56.95% | - | 0.56 |
| | TL3 | 4549 | 7.69 | 20.98% | - | 0.78 |
| | WSJ | 617 | 1.34 | 14.78% | - | 0.44 |
| WCS | AMI | 10071 | 5.30 | 35.34% | - | 0.65 |
| | LSP | 1738 | 4.64 | 38.85% | - | 0.67 |
| | SWB | 22282 | 13.86 | 76.21% | - | 0.55 |
| | TL3 | 9080 | 8.13 | 42.68% | - | 0.68 |
| | WSJ | 49 | 0.07 | 18.31% | - | 0.37 |
| WE & ADS | AMI | 424 | 0.61 | 7.43% | 1.69% | 0.82 |
| | LSP | 2414 | 8.69 | 3.59% | 1.88% | 0.88 |
| | SWB | 520 | 1.11 | 13.67% | 1.72% | 0.79 |
| | TL3 | 9901 | 21.57 | 5.87% | 1.78% | 0.87 |
| | WSJ | 7 | 0.02 | 6.16% | 1.88% | 0.77 |

### 4.5. Automatic Speech Recognition Performance

The results in Table 3 show that the data selected using a random strategy and WCS did not improve ASR performance; the WER increased when 8 or 32 additional hours of selected data were added to the 32-hour supervised baseline. In contrast, the WER for the HuBERT based model on CH5 Test decreased to 22.86% and 22.61% with 8 and 32 hours of selected data, respectively. This WER reduction can be compared with the performance extrapolated from the plot in Figure 2.

Table 3: *Semi-supervised ASR performance on CH5 Test with 8 and 32 hours of data selected from the data pool.*

| Data Selection | Manu. + Auto. (hours) | WER (%) | WER Recovery |
|---|---|---|---|
| Supervision | 32 | 23.17 | |
| (extrapolation) | 40 | 22.24 | 100% |
| random | 32 + 8 | 23.32 | -16.13% |
| WCS | 32 + 8 | 23.48 | -33.33% |
| $\widehat{\text{WER}}$ & ADS | 32 + 8 | 22.86 | 33.33% |
| (extrapolation) | 64 | 21.07 | 100% |
| random | 32 + 32 | 23.21 | -1.90% |
| WCS | 32 + 32 | 23.69 | -24.76% |
| $\widehat{\text{WER}}$ & ADS | 32 + 32 | **22.61** | 26.66% |

## 5. Conclusion

A novel approach to data selection for SSL in ASR is proposed with an untranscribed multi-domain data pool, leveraging WE and ADS. To obtain highly accurate ASR transcripts for training, their WER was estimated using MTR-ER, demonstrating improved generalisation across multiple domains. Simultaneously, to reduce domain mismatch, ADS was defined and measured using noise-contrastive estimation. By combining these two criteria, the data selection method gained 26.66% of the expected performance improvement by fully supervised training.

# 6. References


[1] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty *et al.*, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck *et al.*, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[2] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng *et al.*, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," in *Interspeech 2021*, 2021, pp. 3670–3674.

[3] K. Veselý, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 267–272.

[4] K. Veselý, L. Burget, and J. Černocký, "Semi-supervised DNN training with word selection for ASR," in *Proceedings of Interspeech 2017*, 2017, pp. 3687–3691.

[5] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap *et al.*, "End-to-End ASR: From supervised to semi-supervised learning with modern architectures," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, 2020. [Online]. Available: https://openreview.net/forum?id=OSVxDDc360z

[6] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor *et al.*, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*.

[7] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised dnn training in meeting recognition," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 141–146.

[8] G. Zavaliagkos, M.-H. Siu, T. Colthurst, and J. Billa, "Using untranscribed training data to improve performance," in *Proceedings of 5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998, p. paper 1007.

[9] A.-L. Georgescu, C. Manolache, D. Oneaţă, H. Cucu, and C. Burileanu, "Data-filtering methods for self-training of automatic speech recognition systems," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 1–7.

[10] Q. Li, D. Qiu, Y. Zhang, B. Li, Y. He, P. C. Woodland *et al.*, "Confidence estimation for attention-based sequence-to-sequence models for speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6388–6392.

[11] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna, "Quality estimation for automatic speech recognition," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, J. Tsujii and J. Hajic, Eds. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 1813–1823. [Online]. Available: https://aclanthology.org/C14-1171

[12] J. G. C. de Souza, H. Zamani, M. Negri, M. Turchi, and D. Falavigna, "Multitask learning for adaptive quality estimation of automatically transcribed utterances," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, R. Mihalcea, J. Chai, and A. Sarkar, Eds. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 714–724. [Online]. Available: https://aclanthology.org/N15-1073

[13] S. A. Chowdhury and A. Ali, "Multilingual word error rate estimation: E-Wer3," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] C. Park, C. Lu, M. Chen, and T. Hain, "Fast word error rate estimation using self-supervised representations for speech and text," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[15] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "Survey on multi-output learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2409–2429, 2020.

[16] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: Treating targets as inputs," *Machine Learning*, vol. 104, pp. 55–98, 2016.

[17] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proceedings of Interspeech 2015*, 2015, pp. 2897–2901.

[18] Y. Chen, W. Ding, and J. Lai, "Improving noisy student training on non-target domain data for automatic speech recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[19] C. Park, R. Ahmad, and T. Hain, "Unsupervised data selection for speech recognition with contrastive loss ratios," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8587–8591.

[20] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proceedings of Interspeech 2019*, 2019, pp. 3465–3469.

[21] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 297–304. [Online]. Available: https://proceedings.mlr.press/v9/gutmann10a.html

[22] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4297–4300.

[23] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proceedings of 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, 2018, pp. 35–40.

[24] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI'05. Berlin, Heidelberg: Springer-Verlag, 2005, p. 28–39. [Online]. Available: https://doi.org/10.1007/11677482_3

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

[27] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Cham: Springer International Publishing, 2018, pp. 198–208.

[28] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of Workshop Speech Natural Lang.*, ser. HLT '91. USA: Association for Computational Linguistics, 1992, pp. 357–362, [Online]. doi: 10.3115/1075527.1075614.

[29] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré *et al.*, "Libri-Light: A benchmark for ASR with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.