



Distillation Strategies for Discriminative Speech Recognition Rescoring

Prashanth Gurunath Shivakumar, Jari Kolehmainen, Yile Gu, Ankur Gandhe, Ariya Rastrow and Ivan Bulyko

Amazon Alexa AI, USA

{psshvak, jkolehm, yilegu, aggandhe, arastrow, ibbulenko}@amazon.com

Abstract

Second-pass rescoring is employed in most state-of-the-art speech recognition systems. Recently, BERT based models have gained popularity for re-ranking the n-best hypothesis by exploiting the knowledge from masked language model pre-training. Further, fine-tuning with discriminative loss such as minimum word error rate (MWER) has shown to perform better than likelihood-based loss. Streaming applications with low latency requirements impose significant constraints on the size of the models, thereby limiting the word error rate (WER) performance gains. In this paper, we propose effective strategies for distilling from large models discriminatively trained with the MWER objective. We experiment on Librispeech and production scale internal dataset for voice-assistant. Our results demonstrate relative improvements of upto 7% WER over student models trained with MWER. We also show that the proposed distillation can reduce the WER gap between the student and the teacher by 62% upto 100%.

Index Terms: minimum word error rate, ASR rescoring, distillation, BERT

1. Introduction

Two-pass automatic speech recognition (ASR) systems comprise of a first-pass model to generate n-best hypotheses and a second-pass rescoring model to re-rank and pick the best hypothesis. Directly optimizing the second-pass models with respect to the WER is an appealing proposition. One way to achieve this is through the MWER objective, which is designed to directly minimize the expected WER [1, 2]. There have been several attempts at incorporating discriminative training using MWER loss for LSTM-RNN [1, 3], Sequence-to-Sequence [4, 5, 6], RNN-Transducer [7], Transformer-Transducer [8] based architectures for ASR. Recent studies have shown BERT models [9] with bi-directional information encoding and pre-trained knowledge are favorable for rescoring, both in log-likelihood [10] and discriminative loss [2] settings.

Knowledge distillation in neural networks [11] has shown great success in efficiently compressing and mimicking one or ensemble of larger models (teachers). This enables use of much smaller models with the advantages of reduced training time, inference and latency costs. [11] proposed cross-entropy based distillation based on KL-divergence loss for distilling from an ensemble of teacher models for classification problems. [12] extended knowledge distillation by minimizing L2-norm of the intermediate representations from hidden layers between the teacher and the student. L2-norm based objectives are also used to closely replicate real value predictions of teacher in application to regression problems [13, 14]. There have also been several successful attempts of knowledge distillation in language

modeling domain. Most settings in the language model domain are posed as a classification problem and typically combination of the KL divergence loss with regularization is used. For BERT models, optimizing the cosine distance between the teacher and student embeddings is found to be useful [15]. [10] proposed to distill pseudo log-likelihoods (PLL) from BERT [16] using L2-loss based regression over the classification (CLS) token in application to utterance rescoring for ASR.

Recently, RescoreBERT was proposed [2] to train a BERT based second-pass rescoring model with several discriminative objective functions including MWER. RescoreBERT provided up-to 5.3% relative WER improvement, which was attributed to MWER fine-tuning. The paper also proposed a technique to distill knowledge to a smaller model of 5M parameters for low-latency streaming applications. The distillation involves training the student BERT model to predict the sentence-level PLL from the bigger 170M parameter teacher model before proceeding with MWER training. Although, reducing the model size from 170M to 5M parameters brings average latency improvement of approximately 1350% (relative), the WER improvements are diminished by 61.3% relative to 170M model. This suggests there is a need for better distillation strategies to attain better latency vs. WER trade-off.

The advantages of using discriminative loss functions such as MWER [2, 3] to train language models are clear. Further, the need for small models to enable low-latency streaming applications is necessary. However, efficient techniques to distill from a model trained with discriminative loss such as MWER have not been explored. Moreover, the question on effect of distillation and its ability to retain the discriminative power with respect to MWER trained model is unanswered. In this paper, we propose several techniques to perform distillation from a second-pass ASR rescoring teacher model trained with a discriminative MWER criterion. We devise loss functions to train the student model to mimic the teacher while retaining the discriminative power of the teacher to minimize the expected WER. We demonstrate that distilling from a larger teacher gives better WER in comparison to a MWER trained student model. To the best of our knowledge, this work is the first to explore distillation from a MWER trained model and explore its feasibility in application to second-pass ASR rescoring.

2. RescoreBERT

RescoreBERT is a discriminative second-pass rescorer based on BERT [2]. The architecture comprises of a BERT model with a feed-forward layer on the CLS token embedding. The model consumes a sequence of tokens $X = (x_1, \dots, x_N)$ and outputs a score. Figure 1 shows the schema for training RescoreBERT as proposed in [2]. The training process can be visioned in three

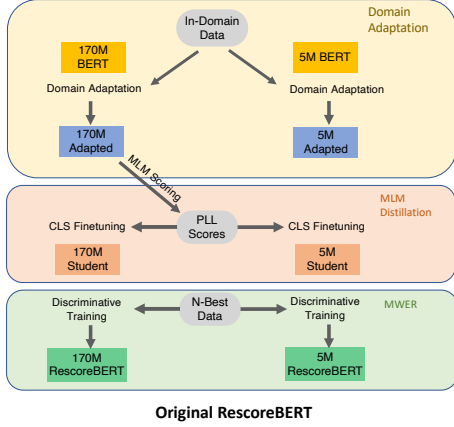


Figure 1: RescoreBERT Training Schema [2]

(boxes represent model training; rounded boxes represent data)

stages: (i) Domain adaptation: pre-trained BERT model is further trained and adapted on the in-domain datasets using the masked language model (MLM) scheme [9], (ii) MLM Distillation: this step involves addition of a feed-forward layer over CLS token embedding and fine-tuning to predict the PLLs [10]:

$$PLL(X) = - \sum_{t=1}^N \log P(x_t | X_{\setminus t}) \quad (1)$$

$$X_{\setminus t} = (\dots, x_{t-1}, [\text{MASK}], x_{t+1}, \dots)$$

and (iii) Discriminative training with MWER loss:

$$L_{MWER} = \sum_{j=1}^n (E_j - \bar{E}) \text{Softmax}(s_j) \quad (2)$$

where E_j is the edit distance of j^{th} hypothesis in an utterance, \bar{E} is the mean edit distance computed over all hypotheses of n -best belonging to the utterance, s_j is the final score of hypothesis j obtained after linear interpolation of the first pass scores with the RescoreBERT score, and the $\text{Softmax}(\cdot)$ gives the posterior distribution of the hypothesis over the n -best for an utterance.

Authors in [2] additionally propose a smaller size model of 5M parameters for low-latency streaming. While most of the training scheme remains identical to the bigger model, the paper proposes to distill the PLLs from the bigger teacher model onto the 5M model during the MLM distillation stage.

3. Proposed Distillation

Study in [2] noted that the low latency 5M model results in diminished WER gains relative to 170M teacher (upto 61.3%). Moreover, we hypothesize that the MLM distillation (where objective is language model scoring) is sub-optimal for the target task of rescoring. In this work, we propose to distill directly from the final MWER-trained teacher model as illustrated in Figure 2. We present several loss functions for distillation to best preserve the discriminative power of MWER models.

3.1. Cross-Entropy over n -best posterior distributions

We propose to compute the cross-entropy between the posterior n -best distribution between the teacher and the student. In order to compute the MWER loss objective (see Equation (2)), we

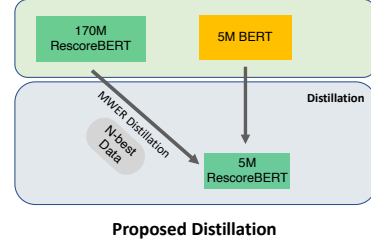


Figure 2: Proposed scheme for distilling from MWER models.

compute the softmax over the hypotheses over the n -best of the utterance. We can directly optimize the student to mimic this distribution from the teacher, by using the following objective:

$$L_{post} = CE(\text{Softmax}(s_j^s/T), \text{Softmax}(s_j^t/T)) \quad (3)$$

where s_j^s is the predicted score from the student for hypothesis j of an utterance, s_j^t is the score from the teacher model for the same hypothesis, T is the temperature parameter suggested in [11], $CE(\cdot)$ is the cross-entropy function and $\text{Softmax}(s_j/T) = \frac{\exp(-s_j/T)}{\sum_{k=1}^n \exp(-s_k/T)}$.

3.2. Oracle based posterior correction

One downside of the previous cross-entropy based distillation is that the upper-bound of the student is limited to that of the teacher. [11] proposed having an additional cross-entropy loss against the groundtruth labels to account for instances when the teacher itself is sub-optimal. However, since there is no ground-truth with PLL scores for MWER objective, we propose to use oracle hypothesis as the ground-truth [17] and add this as a correction term to the previously proposed loss in Equation (3):

$$L_{pOracle} = L_{post} + \alpha * CE(\text{Softmax}(s_j^s), \text{OneHot}(\arg \min_j(E))) \quad (4)$$

where $\text{OneHot}(\cdot)$ is the one-hot encoding function, and α is a tunable parameter. The second term is estimating the oracle multinomial distribution over the n -best, whereas the first term mimics the distribution of the teacher.

3.3. Mean Squared Error Loss

We also explore the typical mean squared error (MSE) loss for distilling the score output directly from the teacher. We experiment with two variations of the MSE loss for distillation. First, where we sample hypothesis by maintaining the n -best structure (i.e., hypotheses belonging to an utterance appear in the same batch during training) and consider the 1st pass scores during optimization (we refer to this as n -best MSE):

$$L_{nMSE} = \frac{1}{n} \sum_{j=1}^n (s_j^t - s_j^s)^2 \quad (5)$$

Second, by random sampling and including the scores from second-pass only (we refer to this as MSE):

$$L_{MSE} = \frac{1}{N} \sum_i^N (\hat{s}_i^t - \hat{s}_i^s)^2 \quad (6)$$

\hat{s} indicates the score is from second-pass only, N is the batch size. This method has the advantage of using additional text-only data during the distillation process.

Dataset	Hours/Tokens
Domain adaptation training set	3.8B tokens
Domain adaptation development set	40M tokens
N-best training set	95K hours
N-best development set	49 hours
General	194 hours
Knowledge/Information Domain	20 hours
Navigation Domain	39 hours
Shopping Domain	5 hours
Music Domain	40 hours
Tail dataset	100 hours

Table 1: *Data splits. Bottom half represent test-sets.*

4. Experiments

4.1. Data

For experimentation, we use the publicly available Librispeech [18] and in-house dataset consisting of de-identified, far-field conversations with voice assistants for English. Our in-house dataset is a compilation of multiple domains including Information/Knowledge, Navigation, Music, Video, Shopping and generic de-identified user interactions. Partitions (non-overlapping) for training, development and testing are listed in Table 1. Note, in the domain adaptation phase, we are not constrained to audio only data which allows us to use magnitudes more data for each of the domains. Audio data is used for MWER training and distillation experiments. For testing purposes, we present the results on aforementioned domain specific datasets (see bottom half of Table 1). The music domain dataset is a compilation of synthetic speech generated using text-to-speech (TTS) system. The general test set is a compilation of de-identified user interactions with AI conversational agent and the tail dataset is a subset sampled from general test set and is curated based on tail linguistic data distribution.

4.2. Experimental Setup

4.2.1. ASR system

In case of Librispeech, we employ pre-trained, publicly available, tiny English Whisper model [19] which is based on transformer sequence-to-sequence architecture, as the first pass. For the internal dataset, the first pass model is based on RNN-T architecture [20] with shallow fusion n-gram language model [21, 22]. The RNN-T model is trained using the audio data listed in Table 1 and also employs data augmentation using semi-supervised learning techniques. The n-best size from the first pass is restricted to 10 for rescoring.

4.2.2. Rescoring Model

Two BERT models differing in model sizes are used: (i) 170M parameter, and (ii) 5M parameter model. The Teacher model comprises 16 layers with hidden size of 1024 and 16 attention heads (total of 170M parameters excluding embedding layer). The student model comprises 4 layers with hidden size of 320 and 16 attention heads (total of 5M parameters excluding embedding layer). Sentence-piece tokenizer trained on multilingual data [23] with a vocabulary size of 153,623 is adopted.

For Librispeech experiments, we use mini-BERT as student, base-BERT as teacher and closely follow the RescoreBERT setup in [2]. In case of internal dataset, the BERT pre-training is performed on MC4 dataset [23]. Further, domain

DataSets	% WERR over 1st pass		
	PLL 5M	Baseline 5M	Teacher 170M
General	-1.15	-3.16	-4.20
Knowledge	-4.21	-6.33	-10.54
Navigation	-2.31	-3.66	-6.78
Shopping	-1.82	-7.03	-10.82
Music	-7.94	-20.45	-23.77
Tail	-0.68	-1.67	-3.01

Table 2: *WER relative improvements over first pass.*

adaptation is performed on in-domain dataset as listed in Table 1. Given larger data size, we found that the MLM distillation did not provide any improvements and hence skip it in this work. Finally, a feed-forward layer is added to the CLS embedding and trained with MWER objective until convergence. Interpolation weights for the first-pass is set to 20.0 and second-pass to 1.0 during training. For run-time inference, the weights are optimized on the development dataset.

For the proposed distillation techniques, the 170M model described under Section 4.2.2 is used as the teacher to score each hypotheses in the n-best data. The scores are distilled to the student whose architecture is identical to the baseline 5M model using losses proposed under Section 3. Parameter α in Equation (4) is set to 1.0 for our experiments.

4.3. Implementation

We use a batch size of 48 during domain adaptation. Batch size of 16 and 64 is used during MWER training for 170M and 5M model respectively. Learning rate decay schedule is adopted with initial learning rate set to $1e-4$ for 5M model and $1e-5$ for the 170M model. Adam optimizer is adopted during training. The training is continued until convergence on the development dataset. The baseline model, teacher and the distilled models are all trained on identical data for fair comparison.

4.4. Baseline and Evaluation

We use two baseline systems in this study: (i) the 5M parameter RescoreBERT model proposed in [2] (see Section 4.2.2), and (ii) the PLL based rescoring model distilled from 170M domain adapted MLM [10]. The PLL baseline is used to verify claims that the proposed distillation technique is able to preserve the discriminative power of MWER trained teacher. The 5M RescoreBERT baseline is used for all other experiments.

Evaluations are performed on unseen held-out test sets listed in Table 1. We present results in terms of percentage word error rate reductions (WERR) over the baseline 5M model. We would like to note that the absolute WER is less than 13% for synthetic datasets and well below 10% for others. In order to assess the effectiveness of each of the distillation techniques, we also provide the percentage improvement relative to WERR from teacher over 5M baseline model.

5. Results

We first perform experiments on the internal datasets and then use the best distillation technique to present results on Librispeech. Table 2 shows the performances of the baseline systems and the 170M RescoreBERT teacher model on internal datasets. Compared to using PLL, the benefits of MWER training is clear. There is substantial performance gap between the baseline 5M and the teacher attributed to bigger model size.

Model	% WERR on test-sets					
	General	Knowledge	Navigation	Shopping	Music [†]	Tail
Teacher 170M	-1.19	-4.5	-3.38	-4.55	-7.72	-1.39
L_{MSE}	-0.40 (33.6)	-1.93 (42.9)	-1.27 (37.6)	-3.79 (83.3)	-5.79 (75.0)	-1.25 (89.9)
L_{nMSE}	-0.79 (66.4)	-2.25 (50.0)	-1.27 (37.6)	-3.90 (85.7)	-5.56 (72.0)	-1.11 (79.9)
L_{post}	-0.99 (83.2)	-3.22 (71.6)	-2.11 (62.4)	-2.60 (57.1)	-5.56 (72.0)	-1.11 (79.9)
$L_{pOracle}$	-0.99 (83.2)	-2.25 (50.0)	-1.13 (33.4)	-3.25 (71.4)	-4.78 (61.9)	-1.11 (79.9)
$\beta L_{nMSE} + L_{MWER}$	-0.79 (66.4)	-2.57 (57.1)	-1.83 (54.1)	-4.22 (92.7)	-5.63 (72.9)	-1.25 (89.9)
$L_{post} + \beta L_{nMSE}$	-1.19 (100.)	-2.24 (50.0)	-1.13 (33.4)	-3.03 (66.6)	-5.17 (67.0)	-1.39 (100.)
$L_{post} + \beta L_{MWER}$	-0.79 (66.4)	-3.85 (85.5)	-1.27 (37.6)	-3.90 (85.7)	-5.09 (65.9)	-1.25 (89.9)
$\gamma L_{post} + \beta L_{nMSE} + \gamma L_{MWER}$	-0.79 (66.4)	-3.21 (71.5)	-1.55 (45.9)	-3.57 (78.5)	-7.02 (90.9)	-1.25 (89.9)

Table 3: Results of rescoring models in terms of word error rate reduction (WERR) relative to the baseline 5M model on different test-sets. †indicates the test-set is synthetic. Numbers inside parenthesis are percentage improvements from distillation relative to WER gap between teacher and baseline. $\beta=0.01; \gamma=0.5$. All the improvements are statistically significant with $p < 0.002$ (Matched Pair Sentence Segment - Word Error [24]) over the baseline.

5.1. Initialization strategy for distillation

Next, we performed experiments using vanilla MSE loss (L_{MSE}) to deduce the optimal initialization point for the student model before distillation. We consider, three candidate initialization points: (i) before domain adaptation, i.e., distilling directly to BERT model, (ii) after domain adaptation and before MWER training, and (iii) after MWER training, i.e., using the baseline 5M RescoreBERT as starting point for distillation. We observed initializing on MWER trained baseline provides no benefits, i.e., performing MWER training before distillation is not beneficial. Moreover, distillation without domain adaptation results in a large performance gap for synthetic datasets (Music). Considering the above, we find domain adapting the BERT model a good starting point for distillation and make this a standard for the rest of our experiments.

5.2. Comparison of distillation strategies

Table 3 presents the relative WER improvements obtained using proposed distillation techniques over the 5M baseline model. We consider the 170M teacher as the upper limit and evaluate performance improvements obtained from distillation relative to reducing the gap between teacher and student. Firstly, we find vanilla MSE loss can provide decent improvements over the MWER trained baseline. The MSE loss is particularly effective on synthetic music dataset. We believe this is because of the peculiar acoustic characteristics of the TTS generated signals, which can reduce reliance on the first pass scores. Note, this is the only loss that doesn't account for the first pass scores explicitly. Inclusion of the first pass scores (L_{nMSE}) improves performance on datasets where the initial room for improvement is small (although we observe slightly smaller improvements on tail). Both the MSE objectives are effective on datasets where the gap in student-teacher performance is large. The cross-entropy (L_{post}) loss over posterior n-best distribution gives further improvements especially on the general, knowledge and

Model	test-clean	test-other
Whisper [19]	5.67	12.89
Teacher (base-BERT)	4.48	10.65
Baseline (mini-BERT)	5.32	12.02
Distilled (mini-BERT)	5.21 (13.1%)	11.90 (8.8%)

Table 4: WER Results on Librispeech. Numbers inside parenthesis are percentage improvements from distillation relative to WER gap between teacher and baseline.

navigation test-sets. Optimum results were obtained with temperature set to 2. We find the oracle correction ($L_{pOracle}$) provides benefits on shopping domain, however fails to improve on others. Additionally, we also experiment with linear combinations of MWER, MSE, cross-entropy loss and find it to be better than distilling with single loss objective.

Overall, proposed distillation helps reduce the WER gap between the 170M and 5M models by up-to 93% relative on domain specific datasets. We observe that the distillation is particularly effective for the general and tail data, recovering 100% of teacher performance. The performance of the teacher highlights the modeling challenge for tail data given minimal improvements. This result shows that the distillation is not sensitive to the underlying data distribution. Finally, our best distilled model (L_{post}) provides 4.2%, 9.03%, 5.83%, 10.11%, 21.33% and 2.74% improvement over first pass on General, Knowledge, Navigation, Shopping, Music and Tail test-sets respectively.

5.3. Results on Librispeech

Table 4 lists results obtained using the best performing distillation scheme on Librispeech. The teacher RescoreBERT gives a relative improvement of 21%/17.4% on test-clean/test-other over the first-pass. The baseline RescoreBERT is able to achieve 6.2%/6.7% WERR on test-clean/test-other. The distillation is able to reduce the teacher-baseline gap by 13.1%/8.8% on test-clean/test-other showcasing the effectiveness of the approach on different domains and different first pass models.

6. Conclusions

In this study, we proposed new techniques to distill from a teacher trained with the MWER criterion. A cross-entropy based distillation loss computed over n-best posterior distribution, an oracle based correction to the cross-entropy loss and the traditional MSE loss were explored and compared to deduce the best approach for distillation. We designed experiments and presented evaluations on publicly available Librispeech and domain specific internal datasets. We showed that distilling from a larger teacher model, discriminatively trained on MWER loss, can outperform typical MWER training for the student. Improvements of up-to 7% on WER was obtained over the MWER trained student model. The distillation is shown to reduce the WER gap between the teacher and student by 62% upto 100%. In the future, we plan to distill from a bigger teacher to maximize the impact of distillation.

7. References

- [1] T. Hori, C. Hori, S. Watanabe, and J. Hershey, "Minimum word error training of long short-term memory recurrent neural network language models for speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 5990–5994.
- [2] L. Xu, Y. Gu, J. Kolehmainen, H. Khan, A. Gandhe, A. Rastrow, A. Stolcke, and I. Bulyko, "Rescorebert: Discriminative speech recognition rescoring with bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6117–6121.
- [3] A. Gandhe and A. Rastrow, "Audio-attention discriminative language model for asr rescoring," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7944–7948.
- [4] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *Proc. IEEE ICASSP*, 2018, pp. 4839–4843.
- [5] B. Li, S.-y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohmaier, and Y. Wu, "Towards fast and accurate streaming end-to-end asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6069–6073.
- [6] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohmaier, Y. Wu *et al.*, "Two-pass end-to-end speech recognition," *Proc. Interspeech 2019*, pp. 2773–2777, 2019.
- [7] J. Guo, G. Tiwari, J. Droppo, M. Van Segbroeck, C.-W. Huang, A. Stolcke, and R. Maas, "Efficient minimum word error rate training of rnn-transducer for end-to-end speech recognition," *Proc. Interspeech 2020*, pp. 2807–2811, 2020.
- [8] Z. Meng, Y. Wu, N. Kanda, L. Lu, X. Chen, G. Ye, E. Sun, J. Li, and Y. Gong, "Minimum word error rate training with language model fusion for end-to-end speech recognition," *arXiv preprint arXiv:2106.02302*, 2021.
- [9] 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*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [10] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2699–2712.
- [11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [12] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.
- [13] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Advances in neural information processing systems*, vol. 30, 2017.
- [14] C. Wang, X. Lan, and Y. Zhang, "Model distillation with knowledge transfer from face classification to alignment and verification," *arXiv preprint arXiv:1709.02929*, 2017.
- [15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [16] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using BERT for speech recognition," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 1081–1093.
- [17] E. Variani, T. Chen, J. Apfel, B. Ramabhadran, S. Lee, and P. Moreno, "Neural oracle search on n-best hypotheses," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7824–7828.
- [18] 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.
- [19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [20] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.
- [21] A. Gourav, L. Liu, A. Gandhe, Y. Gu, G. Lan, X. Huang, S. Kalman, G. Tiwari, D. Filimonov, A. Rastrow *et al.*, "Personalization strategies for end-to-end speech recognition systems," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7348–7352.
- [22] V. Ravi, Y. Gu, A. Gandhe, A. Rastrow, L. Liu, D. Filimonov, S. Novotney, and I. Bulyko, "Improving accuracy of rare words for rnn-transducer through unigram shallow fusion," *arXiv preprint arXiv:2012.00133*, 2020.
- [23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [24] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1989, pp. 532–535.