Interspeech 2022
18-22 September 2022, Incheon, Korea

Pseudo Label Is Better Than Human Label

Dongseong Hwang, Khe Chai Sim, Zhouyuan Huo, Trevor Strohman

Google, U.S.A

{dongseong, khechai, zhhuo, strohman}@google.com

Abstract

State-of-the-art automatic speech recognition (ASR) systems are trained with tens of thousands of hours of labeled speech data. Human transcription is expensive and time consuming. Factors such as the quality and consistency of the transcription can greatly affect the performance of the ASR models trained with these data. In this paper, we show that we can train a strong teacher model to produce high quality pseudo labels by utilizing recent self-supervised and semi-supervised learning techniques. Specifically, we use JUST (Joint Unsupervised/Supervised Training) and iterative noisy student teacher training to train a 600 million parameter bi-directional teacher model. This model achieved 4.0% word error rate (WER) on a voice search task, 11.1% relatively better than a baseline. We further show that by using this strong teacher model to generate high-quality pseudo labels for training, we can achieve 13.6% relative WER reduction (5.9% to 5.1%) for a streaming model compared to using human labels.

Index Terms: speech recognition, pseudo label, semi-supervised, self-supervised

1. Introduction

The success of end-to-end (E2E) speech recognition models [1, 2, 3] are highly dependent on having a large amount of high-quality transcribed speech data. However, it is expensive and difficult to obtain and maintain the high-quality human transcriptions, restricting the development of automatic speech recognition (ASR). To relieve the dependence on large amounts of human-labeled data, self-supervised and semi-supervised learning techniques are heavily explored recently by leveraging on large-scale unlabeled speech data.

Self-supervised learning techniques [4, 5, 6, 7, 8, 9] have been shown to be effective for pre-training ASR models and achieved impressive performance for speech recognition tasks by leveraging large-scale unlabeled speech data. In general, training is done in a 2-stage pretrain-finetune fashion where the models are first pre-trained using self-supervised learning using a large amount of unlabeled data and then fine-tuned using supervised data. For example, it was shown that Wav2vec 2.0 [5] pre-training can greatly reduce the amount of labeled data to just 100 hours and still achieve a reasonable word error rate (WER) performance of 4.0% on the LibriSpeech test-other set [10]. However, this is still worse than the model trained with all the training data. On the other hand, W2v-BERT [6] pre-training can achieve 2.8% WER on the testother set with using the full 960 hours of training data.

Another recent direction of research to leverage large-scale unlabeled data is semi-supervised learning, including noisy student training (NST) [11, 12] and consistency-based methods [13, 14]. NST [12] achieves the state-of-the-art (SoTA) performance on LibriSpeech when there are limited supervised speech data with transcriptions. YouTube model mostly relies on pseudo labels by NST [15]. To improve the performance further, [16, 17] utilize a combination of self-supervised and semi-supervised learning as well as adding a third NST stage after the 2-stage pretrain-finetune scheme. Our previous paper [17] demonstrates that the new approach reduces the requirement of human labels significantly, and matches or improves the SoTA performance with only 3% human labels on the Google voice search task. However, it is still unclear whether a better performance can be achieved when more human labels are available. Otherwise, how much human labels should be mixed with pseudo labels to achieve the best performance?

In this paper, we will describe a multi-stage training strategy to train a strong teacher model to produce high-quality pseudo labels using a combination of self-supervised and semi-supervised learning, which has SoTA performance on voice search. We compare the effectiveness of human labels and pseudo labels by training the same model with corresponding transcriptions respectively. Our results show that the model trained with pseudo labels has significantly better WER of 5.1% on voice search compared to that using human labels (5.9%). This shows that the pseudo labels from a strong teacher have better quality than human labels for large-scale ASR model training. We also demonstrate that once we have a strong teacher model, subsequent training can be done entirely using only the pseudo labels.

The remainder of the paper is organized as follows. Section 2 describes the network architecture and the training methods used to train our strong teacher models for pseudo labeling. Section 3 shows the experimental results on the voice search and medium-form tasks. Section 4 presents some ablation studies and discussions.

2. Methods

2.1. Model Architecture

We consider two types of models, one for the bi-directional teacher model and another for the streaming student model. Both models use the input feature vector of size 528, consisting of 4 contiguous frames of 128-dimension log-mel features [3] sub-sampled by a factor of 3 and a one-hot domain-id vector of size 16. The outputs correspond to $4,096$ word pieces [18].

The architecture and training procedure of the teacher model is similar to Conformer XL model (0.6B) in [19]. The audio encoder has 24 Conformer blocks [20] with model dimension 1024. The convolution kernel size is 15 and the self-attention layer consists of 8 heads with 64 left and right context length. The decoder consists of a 2-layer LSTM [21] label encoder with 2048 units projected down to 640 output units, and a joint network with a single feed-forward layer with 640 units. The total number of weights is 0.6B. The model is trained by minimizing the RNN-T loss [22].

The student model uses cascaded encoders for unifying streaming and non-streaming ASR [23]. The 1st-pass causal

  10.21437/Interspeech.2022-11034

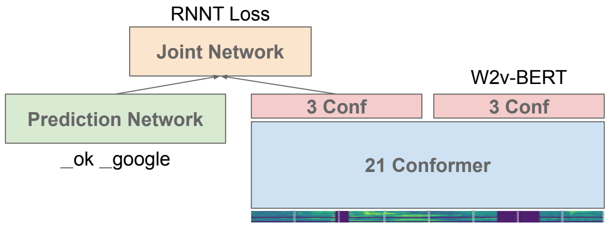

Figure 1: *JUST Hydra – a hydra architecture with separate layers for the last 3 Conformer layers for joint unsupervised and supervised training.*

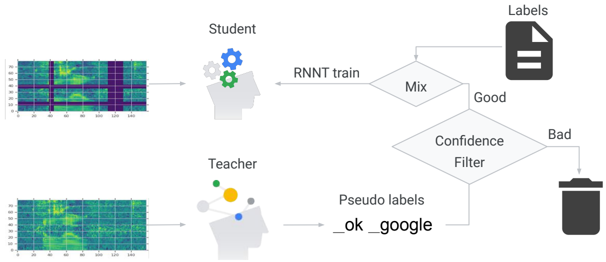

Figure 2: *Noisy student training (NST).*

encoders has 7 Conformer blocks with model dimension 512. The 2nd-pass cascaded encoder has 10 additional non-causal conformer layers that process a total of 900 milliseconds of future audio. Both causal and non-causal encoders feed into a shared hybrid autoregressive transducer (HAT) decoder [24]. This model has a total of number 155M weights.

We use a combination of self-supervised and semi-supervised learning to train a strong teacher model. Specifically, we combine the RNN-T [22] and W2v-BERT [6] losses using a hydra architecture to improve the teacher performance. We further improve the teacher model by applying multiple rounds of noisy student training [11]. We will describe these teacher training improvements in the following sub-sections.

2.2. JUST Hydra

Previous work has shown that combining supervised and self-supervised learning in a multi-task learning fashion can be beneficial [17, 25]. In this work, we use joint unsupervised/supervised training (JUST) [25] to combine the supervised RNNT loss and the self-supervised W2v-BERT loss. JUST simplifies the 2-stage pretrain-finetune scheme into a 1-stage joint training. The representations from the last encoding layers are shared by both the RNNT joint network and W2v-BERT prediction layer. A recent study [26] shows that the last 3 layers in Wav2vec2.0 [5] are specialized to reconstructing masked inputs. This suggests that the last 3 layers should not be shared in JUST. Motivated by this work, we modify JUST to have separate layers for the for the last 3 Conformer layers, while most of lower layers are still shared, as shown in Fig. 1. We call it JUST hydra (JUST with Hydra heads). JUST hydra uses both labeled and unlabeled data. The RNNT loss uses only the labeled data while the W2v-BERT self-supervised loss uses both the labeled and unlabeled data.

2.3. Noisy student training (NST)

In this study, we use noisy student training (NST) [11] in the same way as our previous study [17]. When generating pseudo labels, we use a confidence estimation module (CEM) [27] to filter out low confidence utterances to avoid training with erroneous labels. The filtered pseudo labels are then mixed with human-labeled data to form the training data to learn the student model using the RNNT loss, as shown in Fig. 2.

We combine NST and JUST hydra as follows to progressively improve the teacher model:

1. Use JUST hydra to train a large ASR model using both the labeled and unlabeled data.
2. Using the above model as both the teacher and student, perform self-training to produce the 1st generation NST model.
3. Using the previous NST model as the teacher and initial JUST hydra model as the student, perform self-training to produce the 2nd generation NST.
4. Repeat NST until convergence.

Once we obtain a strong teacher model, the student models are trained in a semi-supervised learning fashion using the pseudo labels generated by the teacher model.

Table 1: *Overview of training data sets.* `MD` *denotes all the supervised data (*`VS` *+* `MF` *+* `en_x` *+* `YT`*, 575k hrs).* $\mathtt{MD}_{\mathtt{unsup}}$ *denotes the unsupervised data (*$\mathtt{VS}_{\mathtt{unsup}}$ *+* $\mathtt{MF}_{\mathtt{unsup}}$*, 1.15M hrs).* `en_x` *includes both* `VS` *and* `MF`*.*

Data set	Label	Hours
Voice search (`VS`)	human	27k
Medium-form (`MF`)	human	26k
Dialect (`en_x`)	human	80k
Youtube (`YT`)	semi	440k
Voice search ($\mathtt{VS}_{\mathtt{unsup}}$)	pseudo	900k
Medium-form ($\mathtt{MF}_{\mathtt{unsup}}$)	pseudo	250k
Voice search ($\mathtt{VS}_{\mathtt{unsup\text{-}new}}$)	pseudo	150k

3. Experiments

3.1. Data

We use a large multi-domain (`MD`) English dataset [28] for training, which consists of utterances from multiple domains, such as voice search (`VS`), medium-form (`MF`) and YouTube (`YT`). All the data are anonymized and hand-transcribed following Google AI principles [29]. `VS` is mostly voice command. `MF` is mostly natural conversation. The `YT` labels are generated from YouTube video transcription with confidence filter [30]. This forms a total of 575k hours of data. In addition, we also use 1.15M hours of unsupervised data for voice search ($\mathtt{VS}_{\mathtt{unsup}}$) and medium-form ($\mathtt{MF}_{\mathtt{unsup}}$). We also used another set of unsupervised voice data ($\mathtt{VS}_{\mathtt{unsup\text{-}new}}$) with 150k hours of data collected at a later time to study the generalization of our teacher models to new data. A summary of the training data is given in Table 1.

JUST hydra uses all the data. NST uses supervised data or all the data. The average utterance lengths for the `VS`, `MF` and `YT` domains are 4.6 seconds, 10.4 seconds and 5 seconds, respectively. For evaluation, we report the WER performance of the voice search (`VS`) and medium-form (`MF`) tasks.

Table 2: *WERs of large models for supervised baseline, JUST hydra and NST.*

Algorithms	Data	WER (%) VS	WER (%) MF
B0 (Supervised)	MD	4.5	2.4
E1 (JUST hydra)	MD + $\mathrm{MD}_{\mathrm{unsup}}$	4.3	2.3
E2 (1st Gen NST, teacher=E1)	MD	4.1	2.1
E3 (2nd Gen NST, teacher=E2)	MD	**4.0**	**2.0**
E4 (3rd Gen NST, teacher=E3)	MD	**4.0**	**2.0**

Table 3: *WERs comparison for streaming models using supervised and NST training.*

Algorithms	WER VS	WER MF
$\mathrm{B0}_{\mathrm{stream}}$ (Supervised)	5.9	3.2
$\mathrm{B1}_{\mathrm{stream}}$ ($\mathrm{B0}_{\mathrm{stream}}$ + EMBR)	5.5	2.9
$\mathrm{E1}_{\mathrm{stream}}$ ($\mathrm{B0}_{\mathrm{stream}}$ + NST)	**5.1**	**2.6**
$\mathrm{E2}_{\mathrm{stream}}$ ($\mathrm{E1}_{\mathrm{stream}}$ + EMBR)	5.5	2.6

Table 4: *WERs for supervised baseline and W2v-BERT learning.*

Algorithms	Data	WER (%) VS	WER (%) MF
B0 (Supervised)	MD	4.5	2.4
E5 (W2v-BERT)	MD + $\mathrm{MD}_{\mathrm{unsup}}$	4.5	2.5
E6 (JUST)	MD + $\mathrm{MD}_{\mathrm{unsup}}$	4.5	2.5
E1 (JUST hydra)	MD + $\mathrm{MD}_{\mathrm{unsup}}$	**4.3**	**2.3**

3.2. Results

3.2.1. *Bi-directional teacher model*

The teacher model is a 0.6B-parameter bi-directional RNNT model. $B0$ in Table 2 is the supervised baseline which is trained with MD using the RNNT loss [31]. E1 is trained using JUST Hydra, where the RNNT loss uses MD, and the W2v-BERT loss uses both MD and $\mathrm{MD}_{\mathrm{unsup}}$. It improves the VS WER from 4.5 to 4.3. On top of E1, we conducted multiple-generation NST to train E2, E3 and E4 using only the MD data with pseudo labels generated by a teacher model. E2 is self-training, where both the teacher and student are based on E1. Subsequent training uses the previous iteration model as teacher and the student model is initialized with E1. It is clear from Table 2 that multiple-generation NST progressively improved the quality of the pseudo labels to yield better teacher models. We obtained the best teacher model after two NST iterations, achieving 4.0 WER on VS and 2.0 on MF. The best performance obtained using human labels is 4.3 and 2.3 on VS and MF, respectively. This shows that using pseudo labels is better than using human labels. In the following section, we will use this teacher model to train production-style streaming models.

3.2.2. *Streaming student model*

The student model is a 155M-parameter cascaded RNNT model. In Table 3, we compare two baseline models trained with MD data using human labels. The $\mathrm{B0}_{\mathrm{stream}}$ baseline trained with the RNNT loss achieved 5.9% and 3.2% WER on VS and MD, respectively. Edit-based minimum Bayes risk (EMBR) [32] further improved the performance to 5.5% and 2.9%. When we perform NST on top of the $\mathrm{B0}_{\mathrm{stream}}$ model, using pseudo labels generated by the strong teacher, E2, from the previous section, we can improve the SoTA performance on VS and MF by a significant margin (5.1% and 2.6%). Note that $\mathrm{B0}_{\mathrm{stream}}$ and $\mathrm{E1}_{\mathrm{stream}}$ have the same architecture and training hyper-parameters. The only difference is that the former is trained with human labels while the latter with pseudo labels. This shows that using pseudo labels is better than using human labels again. We conducted user A/B tests for $\mathrm{B0}_{\mathrm{stream}}$ and $\mathrm{E1}_{\mathrm{stream}}$. $\mathrm{E1}_{\mathrm{stream}}$ won by 30:8 with $< .1\%$ p-value (99.9% confidence).

EMBR training on top of $\mathrm{E1}_{\mathrm{stream}}$ ($\mathrm{E2}_{\mathrm{stream}}$) did not yield further improvement. In fact, we observed degradation on VS.

4. Discussion

4.1. JUST hydra vs W2v-BERT

W2v-BERT [6] for self-supervised learning has been shown to achieve state-of-the-art performance on LibriSpeech [10]. However, we found that the vanilla W2v-BERT (E5) is worse than the baseline (B0), as shown in Table 4. To improve the performance, we used JUST hydra, as described in Section 2.2. Instead of using W2v-BERT self-supervised learning for pre-training, we use JUST [25] to combine RNNT and W2v-BERT losses in a multi-task learning setting. However, using JUST alone (E6) did not improve the performance. With a hydra architecture, we use a separate 3-layer Conformer block to learn the representation for W2v-BERT. As a result, we were able to improve the WER on VS from 4.5 to 4.3 (E1).

As previously described in Section 2.1, the input log-mel features are augmented with a 16-dimensional one-hot domain embedding vector to better handle multiple domains. Since all the feature frames within an utterance are appended with the same domain embedding vector, it does not provide useful information for the W2v-BERT contrastive loss when the negative samples are extracted from the same utterance (used in this paper). When the negative samples come from different utterances within a batch, the domain embedding is too easy to predict. In both cases, the domain information makes W2v-BERT training less effective. JUST hydra (E1) helps to alleviate this problem because jointly training with the RNNT loss can guide the model to learn meaningful domain embedding from data. The Hydra architecture is optimal multi-task architecture fully utilizing both W2v-BERT and RNNT losses.

4.2. Human label vs Pseudo label

In Sections 3.2.1 and 3.2.2, we have shown that using pseudo labels for training is better than using human labels. Here, we investigate whether mixing a small fraction of human labels with pseudo labels could be beneficial. In Table 5, we compared mixing 330, 1k, 2k and 4k hours of human labels when training the teacher model using NST. We observe that using 1k hours of human labels achieve consistent improvements on both the VS and MF tasks. WERs improved from 4.1% to 4.0% on VS, and from 2.1% to 2.0% on MF. The last mixing experiment (E12) uses only human labels, so WERs are same to the student (E1).

On the other hands, mixing either none or 1k hours labels with pseudo labels to train the streaming model achieved

Table 5: *WERs with different amounts of human labels and pseudo labels. All models are trained with* `MD`.

Model	Human labels	WER (%) VS	MF
E2 (Teacher)	All	4.1	2.1
E1 (Student)		4.3	2.3
E8 (NST)	None	4.1	2.1
E9 (NST)	330 hrs	4.1	2.1
E3 (NST, from Table 2)	1k hrs	**4.0**	**2.0**
E10 (NST)	2k hrs	4.1	2.1
E11 (NST)	4k hrs	4.2	2.1
E12 (NST)	All	4.3	2.3

Table 6: *WERs of NST with* 100% *pseudo labels based on various students.*

Model	Data	WER (%) VS	MF
E2 (Teacher)		4.1	2.1
B1 (Supervised)	`YT`	5.3	4.0
B0 (Supervised)	`MD`	4.5	2.4
E1 (JUST hydra)	`MD` + $\texttt{MD}_{\texttt{unsup}}$	4.3	2.3
E13 (NST, teacher=B1)	`MD` w/o labels	**4.1**	**2.1**
E14 (NST, teacher=B0)	`MD` w/o labels	**4.1**	**2.1**
E15 (NST, teacher=E1)	`MD` w/o labels	**4.1**	**2.1**

the same performance of 5.1% and 2.6%, as shown in Table 3 ($\texttt{E1}_{\texttt{stream}}$). Therefore, it is sufficient to rely 100% on pseudo labels once we have a strong teacher model. However, in practice, it may still be necessary to include a small amount of human labeled data. Completely relying on pseudo labels in multi-generation NST leads to deterioration of pseudo label quality as the teacher may reinforce confirmation bias in the pseudo labels.

4.3. Effect of initial student model

We investigated how student models with different WER performance affect the final performance. In Table 6, we compared NST training starting from three different initial student models. `B1` is a poor starting model since it is trained with only `YT` (out-of-domain) data. `B0` is trained using `MD` data with human labels. `E1` is trained with JUST hydra. In these experiments, we used 100% pseudo labels for NST. As shown in Table 6, regardless of the initial student model, they all converged to the same performance of 4.1% and 2.1%. We found that with a strong teacher model, NST is not sensitive the initial student model. As shown in Section 4.1, it is important to apply self-supervised learning to train an initial teacher model. Once we have a high-quality teacher, NST with pseudo labels is sufficient.

4.4. NST on out-of-domain data

So far, we have conducted NST experiments by replacing human labels with pseudo labels for the `MD` (`VS` + `MF` + `en_x` + `YT`). We show the teacher to be able to generate better pseudo labels for the `MD` data. We investigated the pseudo label quality on out-of-domain (OOD) data by the teacher model trained on `MD`. We conducted ablation studies by performing NST on different OOD unsupervised data, as shown in Table 7. We considered two unsupervised voice search datasets ($\texttt{VS}_{\texttt{unsup}}$ and $\texttt{VS}_{\texttt{unsup-new}}$). The former has about 900k hours of data while the latter is a smaller (150k hours) but more recently collected data to capture temporal domain shift. Finally, we also created another `VS + MTR` dataset by artificially corrupting the clean utterances in `VS` using a room simulator, adding varying degrees of noise and reverberation with an average SNR of 12dB [33], a data augmentation technique used for multi-style training (MTR).

Table 7: *WERs of NST with various data for large and production models.*

Data	VS WER (%) Large	Stream
E3 (Teacher)	4.0	
B0, $\text{B0}_{\text{stream}}$ (Student)	4.5	5.9
`VS`	**4.1**	**5.1**
$\texttt{VS}_{\texttt{unsup}}$	**4.1**	5.4
$\texttt{VS}_{\texttt{unsup-new}}$	4.2	5.5
`VS` + `MTR`	**4.1**	5.5

We use NST to train both the large 0.6B-parameter models (`Large`) and 155M-parameter streaming models (`Stream`) with 100% pseudo labels. From Table 7, we observe that NST training for the `Large` model is fairly robust to the unsupervised data used. The `E3` teacher model (trained on `VS`) works well for other OOD data ($\texttt{VS}_{\texttt{unsup}}$, $\texttt{VS}_{\texttt{unsup-new}}$, `VS + MTR`) and achieve similar performance of 4.1–4.2%. On the other hand, the `Stream` models performed slightly worse on OOD unsupervised data (5.4–5.5%). This means that the `Stream` models are sensitive to domain shift. Large bi-directional model is better for unseen domain generalization. Nevertheless, the OOD streaming models are still better than $\text{B0}_{\text{stream}}$, the baseline model trained with human labels (5.9%).

4.5. Limitations and future work

This study demonstrates that a strong teacher can produce high-quality pseudo labels for NST. Once we have a strong teacher, we need only 1k hrs human labels to produce the next generation best teacher model. However, human labels are still needed to train a good initial teacher model, as JUST hydra uses all human labels (`MD`) in Table 4. For languages that do not have sufficient human labeled data, it remains a challenge and open problem to train high quality initial teacher models using as little data as possible.

5. Conclusions

In this paper, we show that by using joint unsupervised and supervised training with W2v-BERT and noisy student training, we can train a strong teacher model that achieved 11.1% relative word error rate improvement compared to a baseline model trained with human labeled data only. This teacher can generate high-quality pseudo labels that are more consistent than the human labels. Semi-supervised learning with the pseudo labels can achieve 13.6% relative word error rate improvement, compared to our best streaming model for the voice search task.

6. References

[1] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-y. Chang, W. Li, R. Alvarez, Z. Chen *et al.*, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6059–6063.

[2] B. Li, A. Gulati, J. Yu, T. N. Sainath, C.-C. Chiu, A. Narayanan, S.-Y. Chang, R. Pang, Y. He, J. Qin *et al.*, "A better and faster end-to-end model for streaming asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5634–5638.

[3] A. Narayanan, R. Prabhavalkar, C.-C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 920–927.

[4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *arXiv preprint arXiv:2108.06209*, 2021.

[7] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, G. Wang, and P. Moreno, "Injecting text in self-supervised speech pretraining," *arXiv preprint arXiv:2108.12226*, 2021.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," *arXiv preprint arXiv:2202.01855*, 2022.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[11] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[12] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *arXiv preprint arXiv:2005.09629*, 2020.

[13] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[14] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[15] T. Doutre, W. Han, M. Ma, Z. Lu, C.-C. Chiu, R. Pang, A. Narayanan, A. Misra, Y. Zhang, and L. Cao, "Improving streaming automatic speech recognition with non-streaming model distillation on unsupervised data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6558–6562.

[16] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2109.13226*, 2021.

[17] D. Hwang, A. Misra, Z. Huo, N. Siddhartha, S. Garg, D. Qiu, K. C. Sim, T. Strohman, F. Beaufays, and Y. He, "Large-scale asr domain adaptation using self-and semi-supervised learning," *arXiv preprint arXiv:2110.00165*, 2021.

[18] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5149–5152.

[19] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[20] A. Gulati, J. Qin, 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 *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 5036–5040. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-3015

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] A. Graves, "Sequence transduction with recurrent neural networks," in *International Conference on Machine Learning: Representation Learning Workshop*, 2012.

[23] A. Narayanan, T. N. Sainath, R. Pang, J. Yu, C.-C. Chiu, R. Prabhavalkar, E. Variani, and T. Strohman, "Cascaded encoders for unifying streaming and non-streaming asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5629–5633.

[24] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (hat)," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6139–6143.

[25] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. C. Sim, and T. N. Sainath, "Joint unsupervised and supervised training for multilingual asr," *arXiv preprint arXiv:2111.08137*, 2021.

[26] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," *arXiv preprint arXiv:2107.04734*, 2021.

[27] D. Qiu, Q. Li, Y. He, Y. Zhang, B. Li, L. Cao, R. Prabhavalkar, D. Bhatia, W. Li, K. Hu *et al.*, "Learning word-level confidence for subword end-to-end asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6393–6397.

[28] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 441–447.

[29] "Artificial intelligence at google: Our principles," https://ai.google/principles/.

[30] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 368–373.

[31] T. Bagby, K. Rao, and K. C. Sim, "Efficient implementation of recurrent neural network transducer in TensorFlow," in *2018 IEEE Spoken Language Technology Workshop*, 2018, pp. 506–512.

[32] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4839–4843.

[33] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," 2017.