INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Semi-supervised end-to-end ASR via teacher-student learning with conditional posterior distribution

Zi-qiang Zhang[1], *Yan Song*[1], *Jian-shu Zhang*[1], *Ian McLoughlin*[1,2], *Li-rong Dai*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China
[2] ICT cluster, Singapore Institute of Technology, Singapore.
zz12375@mail.ustc.edu.cn, {songy, ivm, lrdai}@ustc.edu.cn

Abstract

Encoder-decoder based methods have become popular for automatic speech recognition (ASR), thanks to their simplified processing stages and low reliance on prior knowledge. However, large amounts of acoustic data with paired transcriptions is generally required to train an effective encoder-decoder model, which is expensive, time-consuming to be collected and not always readily available. However unpaired speech data is abundant, hence several semi-supervised learning methods, such as teacher-student (T/S) learning and pseudo-labeling, have recently been proposed to utilize this potentially valuable resource. In this paper, a novel T/S learning with conditional posterior distribution for encoder-decoder based ASR is proposed. Specifically, the 1-best hypotheses and the conditional posterior distribution from the teacher are exploited to provide more effective supervision. Combined with model perturbation techniques, the proposed method reduces WER by 19.2% relatively on the LibriSpeech benchmark, compared with a system trained using only paired data. This outperforms previous reported 1-best hypothesis results on the same task.
Index Terms: semi-supervised learning, ASR, teacher-student learning

1. Introduction

In recent years, attention based encoder-decoder models have shown their ability in many sequence-to-sequence tasks, such as machine translation [1] and automatic speech recognition (ASR) [2, 3]. Despite promising performance, limited availability of paired training data (where both speech records and transcriptions are available) may degrade the effectiveness of encoder-decoder models. However, manual transcription is time-consuming and costly [4], especially for low-resource languages. And thus semi-supervised learning methods, which aim at leveraging large amounts of unpaired data to improve the model's performance, have drawn increasing interest from ASR researchers.

There are several semi-supervised learning based ASR methods in the literature, which will be detailed in Section 2. One typical method is based on teacher-student (T/S) learning [5]. This was first proposed for knowledge distillation where it aims to transfer information from a large highly regularized model into a compact one [5]. In [6, 7], sequence-level T/S learning has been successfully applied to model compression tasks. In the scenario of semi-supervised learning, T/S learning involves first generating supervision information on unlabeled speech from a trained teacher model, and then using this re-paired data to train a student model. The supervision information is usually in the form of an output distribution, making T/S learning different from pseudo-labeling [8], which aims to generate hard labels on unpaired data. Unfortunately, since deriving a sequence-level distribution is intractable due to the huge exploration space involved, it is not easy to apply T/S learning to sequence-to-sequence tasks such as ASR [6, 7]. In [9], 1-best hypothesis approximation based T/S learning was applied for rare word ASR, which is actually similar to pseudo-labeling [10]. Then in [6, 7], 1-best hypothesis and N-best hypotheses were adopted. From [11], it was found that N different 1-best hypotheses obtained from N's decoding with dropout configurations was beneficial. The key point of sequence-level T/S learning is to find an effective approximation of the sequence distribution produced by the teacher model.

In this work, a novel T/S learning method is proposed for encoder-decoder based semi-supervised ASR. Unlike 1-best or N-best hypotheses based T/S learning methods, a new type of supervision for training the student model is presented. Specifically, the teacher model is used to generate two kinds of information concerning label through beam search: 1) 1-best hypothesis just like the previous works [6, 9]; and 2) the sequence of conditional posterior distribution through the 1-best decoding path, which indicates the confidence of the produced symbols during decoding. In the remainder of this paper, the 1-best hypothesis and conditional posterior distribution are named hard- and soft-labels respectively. Exploiting these two levels of information may provide an effective approximation of sequence distribution, and help the student model to tolerate more uncertainty in the unpaired data.

In practice, the effectiveness of the supervision can reduce when the output distributions of the teacher and student models are very close, especially soon after the student model is initialized by the teacher. To address this issue, data augmentation [12] and random dropout are applied during student model training, aiming to introduce some perturbation in the student's output distribution. This idea is motivated by model-consistency loss [13, 14] in semi-supervised image classification, where data and/or model perturbation plays a key role.

We evaluate the proposed T/S learning method on the LibriSpeech [15] corpus, demonstrating a 19.2% relative word error rate (WER) reduction compared to a system using paired data solely. This result also outperforms the previous 1-best hypothesis based T/S learning method [6, 9, 10] under a like-for-like comparison.

2. Related work

Our approach is related to semi-supervised ASR methods which aim to leverage unpaired speech data. These methods can be divided into the following two categories. The first involves

http://dx.doi.org/10.21437/Interspeech.2020-1574

reconstructing speech data, restricting them similar to the real input. It is usually implemented by chaining an ASR model to a reconstruction network, such as a text-to-speech (TTS) system [16, 17, 18], a text-to-encoder (TTE) model [19], or just the decoder of a TTS [20, 21].

The second category, known as pseudo-labeling [8] or self-training [10], involves first generating transcriptions of unpaired speech records using an existing ASR system, and then using the pseudo-paired data to train a new model. T/S learning actually resembles pseudo-labeling when using only 1-hypothesis approximation when applied to ASR task. Pseudo-labeling has been shown useful in training conventional ASR systems [22, 23] in early years. However very recently, an end-to-end model based self-training method with confidence data selection [10] and local prior matching [24] was studied, achieving state-of-the-art WER recovery rate [10] on the LibriSpeech semi-supervised benchmark. We note that a CTC based semi-supervised ASR system [25] also applied data augmentation and dropout to the student model, but that work was based on online pseudo-labeling with hard labels, and without beam search.

3. T/S learning for end-to-end ASR

3.1. End-to-end attention based ASR model

End-to-end ASR models are designed to directly map variable length speech features to a character or word sequence. State-of-the-art end-to-end ASR models usually consist of an encoder and a decoder equipped with an attention network [26, 27]. Input sequence $X = \{x_t|t = 1, \ldots, T'\}$ is first converted to an hidden representation $H = \{h_t|h = 1, \ldots, T\}$ by the encoder, where T' and T are the length of input and representation sequence respectively. The hidden representation is then fed into the decoder through an attention mechanism (Att). The decoder is usually a recurrent network (Rnn), outputting the current state s_i from the attention result c_i,

$$c_i = Att(s_{i-1}, H) \tag{1}$$

$$s_i = Rnn(s_{i-1}, y^*_{i-1}, c_i) \tag{2}$$

where y^*_{i-1} is the previous predicted symbol. The i-th output probability is obtained through the softmax of a linear transformation of s_i,

$$p(y_i|y^*_{1:i-1}, X) = Softmax(linear(s_i)) \tag{3}$$

For the next step, y^*_i is sampled from $p(y_i|y^*_{1:i-1}, X)$ to compute the state s_{i+1} in the subsequent iteration of eqn. 2.

In the training phase, a teacher-forcing technique is applied to replace the previous output symbol y^*_{i-1} with the target label $\overline{y}_{i-1}$, in which the probability of the whole sequence $\boldsymbol{y} = \{y_i|i = 1, \ldots, L\}$ is

$$p(\boldsymbol{y}|X) = \prod_{i=1}^{L} p(y_i|\overline{y}_{1:i-1}, X) \tag{4}$$

then the cross-entropy (CE) loss is computed at the output of the network,

$$\begin{aligned} \mathcal{L}_{CE} &= -\frac{1}{L}\log\left(p(\boldsymbol{y} = \overline{\boldsymbol{y}}|X)\right) \\ &= -\frac{1}{L}\sum_{i=1}^{L}\log\left(p(y_i = \overline{y}_i|\overline{y}_{1:i-1}, X)\right) \end{aligned} \tag{5}$$

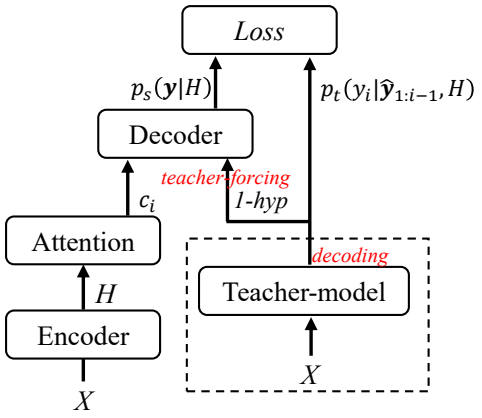

Figure 1: *Model architecture and training process. Two kinds of labels are generated by the teacher model to train the student model: 1-best hypotheses and conditional posterior distributions through the 1-best decoding path.*

Note that the target label sequence $\overline{\boldsymbol{y}} = \{\overline{y}_i|i = 1, \ldots, L\}$ is deterministic, the CE loss is used to maximize the log probability of the corresponding symbol at every time-step i.

3.2. Proposed T/S learning via 1-best hypothesis and conditional posterior distribution

Our proposed T/S learning approach involves three stages. Fig. 1 provides a conceptual illustration.

In the first stage, a teacher model is trained with paired data by supervised learning described in Section 3.1. Then in the second stage, the teacher model is used to generate hard and soft labels for all unlabeled speech samples using a beam search algorithm. Fig. 2 illustrates an example of this process across two sequential time steps. At inference time-step i, k prefix-paths have already been sampled from time-step 1 to $i-1$ according to their decoding scores (log posterior probabilities). Based on each possible prefix-path $\hat{\boldsymbol{y}}_{1:i-1}$, the distribution of the next symbol $p_t(y_i|\hat{\boldsymbol{y}}_{1:i-1}, X)$ can be computed, where $y_i \in \mathcal{Z}$, and $\mathcal{Z}$ is the set of all possible output symbols. Subscript t indicates the teacher model.

Pruning is then executed on each distribution to keep k most likely elements, whose correlative symbols are sampled as the suffix, obtaining $k \times k$ paths (k red boxes in the second column in Fig. 2, where each box corresponds to k paths given one prefix). After adding the log probabilities of the chosen symbols to the decoding scores and comparing them, k best final paths are then selected from the $k \times k$ choices. When beam search proceeds along time, for each prefix $\hat{\boldsymbol{y}}_{1:i-1}$, we record $p_t(y_i|\hat{\boldsymbol{y}}_{1:i-1}, X)$ of all possible y_i, called conditional posterior distributions here.

After this procedure completes, we can easily pick the best possible path (1-best hypothesis), $\hat{\boldsymbol{y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\}$, and the conditional posterior distribution sequence through the path,

$$\{p_t\left(y_i|\hat{\boldsymbol{y}}_{1:i-1}, X\right)|i = 1, \ldots, L\} \tag{6}$$

where $\hat{\boldsymbol{y}}_{1:,i-1}$ is the prefix of $\hat{y}_i$.

The final stage is to train the student model. Following [6], we define the sequence-level T/S learning loss as,

$$\mathcal{L}_{T-S} = -\frac{1}{L}\sum_{\boldsymbol{y}\in\mathcal{Y}} p_t(\boldsymbol{y}|X)\log\left(p_s(\boldsymbol{y}|X)\right) \tag{7}$$

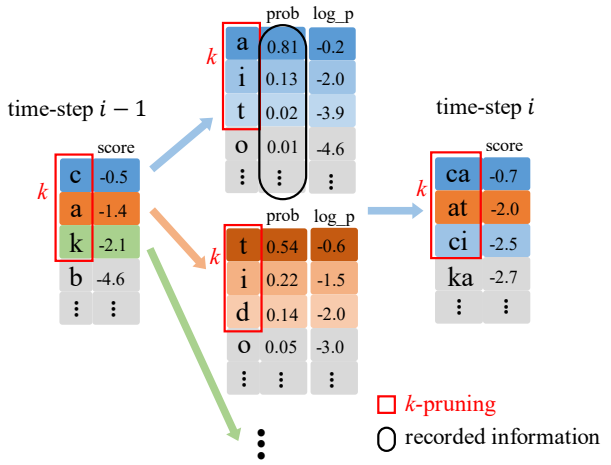

Figure 2: *An example of generating hard and soft labels. Assume the final 1-best hypothesis is **...ca...**, then the conditional posterior distribution recorded at time-step i should be the vector in the black ellipse.*

where subscript s indicates the student model. $\mathcal{Y}$ is the set of all possible paths, which is intractable. Using only 1-best hypothesis $\hat{\boldsymbol{y}}$ is an appropriate approximate of eqn. 7 [6, 9] so,

$$\begin{aligned}\mathcal{L}_{T-S} &\approx -\frac{1}{L}\log\left(p_s(\boldsymbol{y}=\hat{\boldsymbol{y}}|X)\right)\\ &= -\frac{1}{L}\sum_{i=1}^{L}\log\left(p_s(y_i=\hat{y}_i|\hat{y}_{1:i-1},X)\right)\end{aligned} \tag{8}$$

Note that eqn. 8 would be the same as eqn. 5, if we replaced $\hat{\boldsymbol{y}}$ with the true label $\overline{\boldsymbol{y}}$, maximizing the log probability of symbol $\hat{y}_i$ at every time-step. But here we suggest a further improvement. Specifically, the 1-best hypothesis is used for teacher forcing as usual, but at time-step i we do not maximize the exact probability of the chosen symbol $\hat{y}_i$, instead we minimize the cross entropy between the distribution of y_i and the conditional posterior distribution generated from the teacher,

$$\mathcal{L}_{T-S} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{z\in\mathcal{Z}} p_t(z|\hat{\boldsymbol{y}}_{1:i-1},X)\log\left(p_s(z|\hat{\boldsymbol{y}}_{1:i-1},X)\right) \tag{9}$$

In the case that the student is initialized by the teacher, the output distributions of these two models are very similar, making eqn. 9 less effective. To address this issue, we consider introducing a perturbation to the student output distribution. In this way the supervision information from the teacher is generally more accurate than the output of the student, and is better able to guide the student's training. Several techniques to introduce random perturbation could be chosen at this point, and here we apply spectrum data augmentation [12] and dropout.

4. Experimental setup

In this section, we construct experiments to assess the basic aims of the proposed T/S learning method, namely; (1) that T/S learning can help improve the performance of ASR by leveraging additional unlabeled speech data; (2) that using conditional posterior distribution (eqn. 9) can further improve ASR performance.

4.1. Dataset

Our experiments were conducted on the Wall Street Journal (WSJ) [28] and LibriSpeech [15] corpora. We used different subsets as paired and unpaired training set, following the setting of recent works [17, 10]. Specifically, for WSJ the whole training set *si284* contains about 81 hours of speech records, including a 15-hour subset *si84*. The latter was used as paired data while the remainder of *si284* was used as unpaired data. For LibriSpeech, two clean speech subsets, *train_clean_100* (100 hours) and *train_clean_360* (360 hours), were used as paired data and unpaired data respectively. Validation sets for WSJ and LibriSpeech experiments are the same as [17, 10].

4.2. Model details

The encoder-decoder architectures of the student and teacher models were the same. 83 dimensional input features comprised 80 filter-bank and 3 pitch coefficients. For WSJ experiments, the encoder stacked 4 convolutional layers with 64, 64, 128, 128 channels, and 4 BiLSTM [29] layers each with 800 units, the decoder contained one LSTM layer with 800 units, connected to the encoder by location aware attention [30]. 2×2 pooling was performed after every two convolutional layers to downsample the input sequence. For LibriSpeech experiments, we adopted the same architecture as the LAS-medium model [31], in which the encoder stacked two 3×3 convolutional layers with a stride of 2, four BiLSTM layers with projection of 1024 units, and a decoder containing two LSTM layers with 512 units. Character level symbols were modeled in all of our experiments. Both character error rate (CER) and word error rate (WER) were used for evaluation. AdaDelta [32] was adopted to optimize the training, with the epsilon decay and early stop based on the character error rate over the validation set. Training batch size was set to 30. The decoding beam for WSJ experiments was 30, and for LibriSpeech was 10, both for generating labels on unlabeled data and reporting the performance of the test set. No language model (LM) fusion was used through all of our experiments and networks were implemented based on ESPnet [33].

4.3. Training procedure

A teacher model was first trained by standard supervised learning. Then we used the trained teacher model to generate 1-best hypotheses and conditional posterior distributions for the whole paired & unpaired set. For student model training, our proposed training loss (eqn. 9) was used, and eqn. 8 also performed for comparison. Student models started from random initialization for WSJ experiments, while they were initialized by the teacher model for LibriSpeech experiments.

4.4. Data selection & augmentation

We considered applying a data-selection mechanism to filter out some bad-transcribed data, which is also used in [23, 11, 10]. Specifically, we empirically filtered out 20% of the data according to their decoding scores. Initial experiments showed that it did not improve performance on the LibriSpeech experiments, hence data selection was not applied there.

When training the student model on LibriSpeech, we applied spectrum data augmentation [12] (SpecAugment), specifically, the LD policy in [12]. Dropout was also applied after every projection layer of encoders, with a drop probability of 0.2. Considering the influence of the inherent regularization effect of SpecAugment and dropout, we also trained the supervised baseline with these two techniques.

5. Results and analysis

In this section we present experiments on both WSJ and LibriSpeech. Baseline models trained by supervised learning with paired data are denoted Teacher (T) or Oracle (O), according to the data amount they used, with several variants of student system evaluated to identify separately the effect of difference proposed enhancements.

Table 1: *Performance of ASR models on WSJ eval92 via T/S learning. The result of each student model is the average of 5 runs with different random initialization.*

Model	Train. set	Train. label	Data Sel.	Test CER	Test WER
Teacher (T)	*si84*	transcript	-	9.8	26.4
Student I	*si284*	1-hyp	no	8.7	24.6
Student II	*si284*	1-hyp	yes	8.4	24.1
Student III	*si284*	1-hyp+prob	yes	8.2	23.0
Oracle (O)	*si284*	transcript	-	4.4	12.6

Table 1 presents T/S learning systems that have been trained and evaluated on the WSJ corpus, all without SpecAugment or dropout. We first note that the large performance difference between the T and O systems provides a clear indication of the benefit to the latter of the additional labeled data (around 3-5$\times$).

When we then train a straightforward T/S learning system [6, 9] (eqn. 8) with that additional unpaired data (Student I), we note a small improvement in WER (of around 6.8%) to 24.6%. Applying data selection (Student II), which means removing some bad samples which can bring errors when training, further reduces WER slightly to 24.1%. Now incorporating our proposed training loss (eqn. 9), in Student III, ASR performance improves more – reducing CER and WER to 8.2% and 23.0% ;relative improvements of 16.3% and 12.8% respectively.

We believe that such improvement is gained from replacing hard targets from the 1-best hypotheses with the conditional posterior distributions because it allows more information to be provided to the student model. This helps the student to generalize better, as well as handle occasional erroneous teacher assignments in a more advantageous way.

Table 2: *Results tested on LibriSpeech test_clean.*

Model	Train set	Specaug. Drop.	Train label	Test CER	Test WER
T1	LS-100	$\times$	transcript	7.0	16.4
T2	LS-100	✓	transcript	4.3	10.4
S1	LS-460	$\times$	1-hyp	6.2	14.9
S2	LS-460	$\times$	1-hyp+prob	6.0	14.4
S3	LS-460	✓	1-hyp	3.6	9.1
S4	LS-460	✓	1-hyp+prob	3.3	8.4
O1	LS-460	$\times$	transcript	3.3	8.5
O2	LS-460	✓	transcript	2.3	6.2

The results of the LibriSpeech experiments are now presented in Table 2 for several system variants. Again we see a very large performance difference between baseline teacher and oracle models (T1 and O1 respectively). Also as we saw for the WSJ results, the performance gain on the LibriSpeech corpus of using the proposed loss (eqn. 9), in system S2 compared to the previous method (S1), reduces CER and WER slightly, to 6.0% and 14.4% (which is a relative improvement over the teacher of 14.3% and 12.2% respectively).

Incorporating data augmentation and dropout clearly provides performance gains to all systems. However when this is used to introduce a perturbation to the student model (S4 vs. S2 and S3 vs. S1), ASR performance improves markedly. In fact, it is remarkable to note that the student model trained by our proposed method with SpecAugment and dropout (S4), can match the 3.3% CER performance of the baseline oracle model (without Specaug./Drop., O1).

If we consider these results in terms of relative WER reduction, SpecAugment and Dropout together improve the teacher by 36.6% (T1$\rightarrow$T2) and the oracle by 27.1% (O1$\rightarrow$O2), thanks to the inherent regularization effect. However the student is improved by 41.7% (S2$\rightarrow$S4). This greater degree of improvement is thanks to the perturbation effect.

Table 3: *Comparison of the literature of ASR systems trained with LibriSpeech 100h paired data + 360h unpaired data.*

Model	WER	WER $\downarrow$	WRR
Cycle ASR$\rightarrow$TTE [19]	21.5	14.7%	27.6%
Cycle ASR$\leftrightarrow$TTS [17]	17.5	16.7%	38.0%
Pseudo-labeling [24]	12.57	15.3%	33.2%
Pseudo-labeling (ours, S3)	9.1	12.5%	30.9%
Our proposed system	8.4	19.2%	47.6%

Table 3 further presents results reported in several recent works. All WERs reported here were obtained on *test_clean* without language model (LM) fusion. Note that S3 in Table 2 can be seen as a re-implementation of Pseudo-labeling, without well-designed data selection technique in [10]. The relative WER reduction (WER $\downarrow$) and WER recovery rate (defined in [10]) show that our proposed T/S learning surpasses the existing semi-supervised end-to-end ASR methods. It is worth noting at this point that pseudo label quality can be improved by incorporating a LM [10, 24]. For example, the WER in line 3 of Table 3 can be reduced from 12.57% to 9.62% when fusing with a strong LM to refine better hypotheses [10]. Although not demonstrated in this paper, integrating a LM would also be beneficial to our T/S learning method.

6. Conclusion

In this paper we have investigated T/S learning for semi-supervised end-to-end ASR on an encoder-decoder based model. Instead of generating 1-best hypotheses as targets of unlabeled speech, or using N-best hypotheses, we proposed a novel loss function that makes use of the conditional posterior distribution through a 1-best decoding path. Our proposed method incurs minimum modification to the standard supervised training procedure, and can be easily integrated into existing semi-supervised end-to-end frameworks. Experiments on WSJ and the LibriSpeech corpus demonstrate that our proposed method can outperform previous T/S leaning methods, and provides a new perspective for semi-supervised end-to-end ASR.

7. Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Grant No. U1613211), National Key R&D Program of China (2017YFB1002202), the Leading Plan of CAS (XDC08010200), and the Key Science&Technology Project of Anhui Provinces (18030901016).

8. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[4] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, and M. Delcroix, "Semi-supervised end-to-end speech recognition," in *Proc. Interspeech 2018*, 2018, pp. 2–6.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[6] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," *arXiv preprint arXiv:1606.07947*, 2016.

[7] R. M. Mun'im, N. Inoue, and K. Shinoda, "Sequence-level knowledge distillation for model compression of attention-based sequence-to-sequence speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6151–6155.

[8] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Challenges in Representation Learning*, 2013.

[9] B. Li, T. N. Sainath, R. Pang, and Z. Wu, "Semi-supervised training for end-to-end models via weak distillation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2837–2841.

[10] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7084–7088.

[11] S. Dey, P. Motlicek, T. Bui, and F. Dernoncourt, "Exploiting semi-supervised training through a dropout regularization in end-to-end speech recognition," *Proc. Interspeech 2019*, pp. 734–738, 2019.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[13] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems*, 2018, pp. 3235–3246.

[14] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[15] 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)*. IEEE, 2015, pp. 5206–5210.

[16] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 301–308.

[17] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, and J. Černockỳ, "Semi-supervised sequence-to-sequence asr using unpaired speech and text," *Proc. Interspeech 2019*, pp. 3790–3794, 2019.

[18] A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end feedback loss in speech chain framework via straight-through estimator," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6281–6285.

[19] T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux, "Cycle-consistency training for end-to-end speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6271–6275.

[20] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, and T. Nakatani, "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6166–6170.

[21] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," *arXiv preprint arXiv:1905.06791*, 2019.

[22] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for lvcsr," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[23] O. Kapralova, J. Alex, E. Weinstein, P. J. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," in *Fifteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2014.

[24] W.-N. Hsu, A. Lee, G. Synnaeve, and A. Hannun, "Semi-supervised speech recognition via local prior matching," *arXiv preprint arXiv:2002.10336*, 2020.

[25] Y. Chen, W. Wang, and C. Wang, "Semi-supervised asr by end-to-end self-training," *arXiv preprint arXiv:2001.09128*, 2020.

[26] M. K. Baskar, L. Burget, S. Watanabe, M. Karafiát, T. Hori, and J. H. Černockỳ, "Promising accurate prefix boosting for sequence-to-sequence ASR," in *ICASSP*. IEEE, 2019, pp. 5646–5650.

[27] S. Sabour, W. Chan, and M. Norouzi, "Optimal completion distillation for sequence learning," *arXiv preprint arXiv:1810.01398*, 2018.

[28] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

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

[30] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[31] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, and P. Nguyen, "On the choice of modeling unit for sequence-to-sequence speech recognition," *Proc. Interspeech 2019*, pp. 3800–3804, 2019.

[32] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[33] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech 2018*, 2018, pp. 2207–2211.