Leveraging Unlabeled Speech for Sequence Discriminative Training of Acoustic Models

Ashtosh Sapru, Sri Garimella

Amazon Alexa, Bangalore, India

sapru@amazon.com, srigar@amazon.com

Abstract

State-of-the-art Acoustic Modeling (AM) techniques use long short term memory (LSTM) networks, and apply multiple phases of training on large amount of labeled acoustic data - initial cross-entropy (CE) training or connectionist temporal classification (CTC) training followed by sequence discriminative training, such as state-level Minimum Bayes Risk (sMBR). Recently, there is considerable interest in applying Semi-Supervised Learning (SSL) methods that leverage substantial amount of unlabeled speech for improving AM. This paper proposes a novel Teacher-Student based knowledge distillation (KD) approach for sequence discriminative training, where reference state sequence of unlabeled data are estimated using a strong Bi-directional LSTM Teacher model which is then used to guide the sMBR training of a LSTM Student model. We build a strong supervised LSTM AM baseline by using 45000 hours of labeled multi-dialect English data for initial CE or CTC training stage, and 11000 hours of its British English subset for sMBR training phase. To demonstrate the efficacy of the proposed approach, we leverage an additional 38000 hours of unlabeled British English data at only sMBR stage, which yields a relative Word Error Rate (WER) improvement in the range of $6\% - 11\%$ over supervised baselines in clean and noisy test conditions.

Index Terms: Automatic Speech Recognition, Semi-Supervised Learning, Connectionist Temporal Classification, sMBR, Unlabeled Data

1. Introduction

State-of-the-art acoustic models (AM) are large, complex deep neural networks, such as long short term memory (LSTM) networks [1], that typically comprise millions of model parameters. Neural networks can express highly complex input-output relationships and transformations, but the key to get the best performance out of them is the availability of large amounts of labeled acoustic data. For example, [2] reported AM experiments involving 40000 hours of labeled data, in [3] 18000 hours were used for training and 10000 hours were used in [4]. In general, AM training consists of multiple stages: (a) first stage optimizes frame-level Cross-Entropy (CE) loss or sequence-level Connectionist Temporal Classification (CTC) cost [5] or even sequence discriminative criterion such as Lattice-Free Maximum Mutual Information (LF-MMI) [6] (b) final stage minimizes sequence discriminative loss such as state-level Minimum Bayes Risk (sMBR) on word lattices [7]. It is to be noted that having a last stage sMBR is beneficial even when first stage employs a sequence-level loss such as CTC or LF-MMI [8, 9].

Since it is both time consuming and prohibitively expensive to manually transcribe large amounts of acoustic data for every desired condition, Semi-Supervised Learning (SSL) methods were explored, which leverage abundant unlabeled data to further improve accuracies of Automatic Speech Recognition (ASR) systems. Most of SSL methods rely on having access to sufficient amount of transcribed data for building initial seed models, which are then used to machine-label large quantities of unlabeled data. In self-training based SSL methods, a seed ASR model is used to decode unlabeled data and subsequently reliable hypotheses (machine-labelled data) are selected based on confidence scores for ASR training, including Language Models (LM) training [10, 11].

Recently, Knowledge Distillation (KD) techniques based on Teacher-Student (TS) paradigm have become popular for first stage AM training. TS is an effective model compression technique that has been used to bridge the performance gap between larger teacher and smaller student for both frame level CE trained models [12, 13, 14] and CTC trained models [15, 16]. Other works have also explored sequence discriminative training using KD [17, 18]. However, most of these studies have used KD to improve student model on labeled data itself. More recently, KD techniques levaraging unlabeled data at scale have been explored. [19] used unlabeled data and weak distillation to improve tail word recognition accuracy. In [20], about 1 million hours of unlabeled data was used to improve CE trained model. However, subsequent sMBR training was done on labeled data alone. In [17], TS based LF-MMI training was proposed for unsupervised domain adaptation.

Although last stage sMBR training using labeled data has been shown to significantly improve ASR accuracy, to the best of our knowledge, SSL methods for sMBR and large scale training aspects are not well explored in literature. This paper attempts to fill this gap. Specifically, we propose a novel SSL technique based on TS paradigm for improving sMBR using unlabeled data. In addition, we conduct large scale ASR experiments using 38000 hours of unlabeled data on top of 45000 hours of multi-dialect labeled data to demonstrate the efficacy of our technique, for both CE and CTC based LSTM AMs.

The rest of paper is organized as follows. Section 2 presents the Teacher-Student based knowledge distillation (KD) approach used in this study for semi-supervised sMBR training of CE and CTC acoustic models. Section 3 describes unlabeled data selection for SSL based sMBR training, experimental setup used in this study and results. Finally, Section 4 concludes our work and suggests future directions.

2. Teacher Student Based Knowledge Distillation for sMBR

We describe proposed Teacher-Student based KD technique for sMBR. Prior to sMBR training, LSTM AMs are trained using either CE or CTC training criterion. The CE cost function is obtained by accumulating per-frame CE cost over all frames within an utterance and across all utterances. In comparison, CTC loss is obtained by accumulating utterance-level CTC cost,

which is a sequence based cost and involves summation over all possible input output alignments within an utterance corresponding to a ground truth text, across all utterances. Unlike CE or CTC criterion, sMBR criterion incorporates a risk function to maximize expected accuracy of state sequences in a lattice. Formally, a path $\boldsymbol{\pi}$ in the decoding lattice is associated with a label pair $(\mathbf{S}, \mathbf{W})$, where $\mathbf{S}$ is the HMM state sequence corresponding to the word sequence $\mathbf{W}$. For a given feature sequence $\mathbf{X}$, probability of path $\boldsymbol{\pi}$ is dependent on the acoustic and language model scores and is be defined as:

$$P(\boldsymbol{\pi}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{S})^{\kappa} P(\mathbf{W})}{\sum_{\boldsymbol{\pi}} P(\mathbf{X}|\mathbf{S})^{\kappa} P(\mathbf{W})} \quad (1)$$

where, κ is the acoustic scale used in lattice generation. The sMBR objective function can be defined as :

$$\mathcal{F} = \sum_{u} \sum_{\boldsymbol{\pi}} P(\boldsymbol{\pi}|\mathbf{X}_u, \theta)\, A(\mathbf{S}, \mathbf{S}_u) \quad (2)$$

where $\boldsymbol{\theta}$ is the set of model parameters and u is the index of utterance in training set. Each utterance u is associated with a feature sequence $\mathbf{X}_u$, a ground truth $\mathbf{W}_u$ and a sequence of reference HMM states $\mathbf{S}_u$ obtained by force aligning $\mathbf{W}_u$. sMBR is computed as an expectation over paths in the decoding lattice. $A(\mathbf{S}, \mathbf{S}_u)$ captures the raw accuracy, i.e. number of correct state labels, between $\mathbf{S}$ and $\mathbf{S}_u$ while traversing path $\boldsymbol{\pi}$. The gradient computation of sMBR objective is described in [21].

If sMBR training is done on labeled data, reference states are generated by force alignments of the ground truth. However, for unlabeled data reference is not readily available due to lack of ground truth transcription. In this work, we have taken an approach based on knowledge distillation (KD), where reference state label sequence of unlabeled data are estimated using a teacher AM; subsequently, they are used to guide the sMBR training of a student AM. The teacher and student models can have different architectures. We denote the parameters of teacher AM using $\boldsymbol{\theta}_T$. For an utterance u, a path in the decoding lattice of teacher model is represented by $\tilde{\boldsymbol{\pi}}$, with $(\tilde{\mathbf{S}}, \tilde{\mathbf{W}})$ as the associated state and word sequence pair corresponding to $\tilde{\boldsymbol{\pi}}$. Then, KD-sMBR objective is defined as:

$$\mathcal{F}^{KD} = \sum_{u} \sum_{\tilde{\boldsymbol{\pi}}} P(\tilde{\boldsymbol{\pi}}|\mathbf{X}_u, \boldsymbol{\theta}_T) \sum_{\boldsymbol{\pi}} P(\boldsymbol{\pi}|\mathbf{X}_u, \boldsymbol{\theta}) A(\mathbf{S}, \tilde{\mathbf{S}}) \quad (3)$$

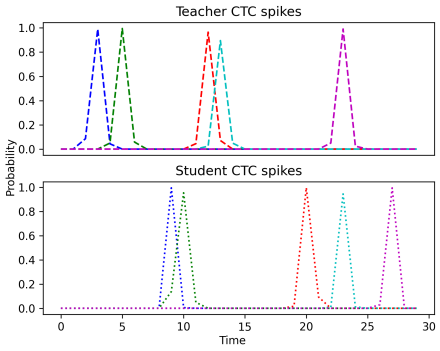
We can approximate Equation 3 by distilling from the teacher model's most probable path. There is evidence to support that this approximation does not compromise the quality of distillation from teacher [22, 23]. If best path in teacher lattice is $\boldsymbol{\pi}_u^T = (\mathbf{S}_u^T, \mathbf{W}_u^T)$ then $P(\tilde{\boldsymbol{\pi}}|\mathbf{X}_u, \boldsymbol{\theta}_T) \approx \delta(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_u^T)$. With this approximation KD-sMBR objective simplifies as:

$$\mathcal{F}^{KD} = \sum_{u} \sum_{\pi} P(\boldsymbol{\pi}|\mathbf{X}_u, \boldsymbol{\theta})\, A(\mathbf{S}, \mathbf{S}_u^T) \quad (4)$$

Although it is possible to obtain teacher model's best hypothesis using a strong LM and subsequently convert it to reference state label sequence for sMBR training, it may not be feasible to scale the data regime to several thousands hours of unlabeled data for SSL sMBR training as decoding may incur lot of computation cost. In addition, since decoding lattices in sMBR training are generated using a weak LM, the reference alignments obtained using a strong LM and teacher AM may not be present in student's decoding lattices. Consequently, the accuracy function may be unable to effectively weigh various lattice paths for sMBR training. Therefore, we propose to use a weak LM and also enforce a constraint that path corresponding to reference alignments estimated using the teacher is present in student's decoding lattice. Summary of KD-sMBR approach for CE initialized model:

1. Generate lattices for the student model using a weak LM. This determines the shared teacher student hypotheses for searching reference state sequence.
2. Rescore lattices generated in step 1 using teacher AM.
3. Compute the best path through the rescored lattice in Step 2., which outputs the reference state label sequence $\mathbf{S}_u^T$.
4. Teacher reference generated in Step 3., is used to drive sMBR training in Equation 4.

Figure 1: *Spike in posteriors of BiLSTM Teacher CTC and LSTM Student CTC occurs at different time frames.*



For CTC initialized models, KD-sMBR algorithm needs to be modified. This is because in CTC trained models, repetitive patterns of targets over consecutive time frames are consumed by the blank symbol. Consequently CTC models produces sparse and spiked target posteriors. However, due to inherent differences in unidirectional LSTM and BiLSTM these posterior spikes often occur at different time frames as shown in Figure 1. In Figure 1, we plot the posteriors obtained by teacher model and student model for a given utterance. It is to be noticed that spikes from teacher and student models are not aligned in time. In fact, the student spikes for the same target is delayed by several frames. Similar observations were also reported in [24]. Sparse posterior spikes observed in CTC models combined with temporal mismatch results in $\mathbf{S}_u^T$ not being in student decoding lattice. However, when both teacher and student models share the same weak LM, it is likely that $\mathbf{W}_u^T$ occurs in some path in student lattice. The state label sequence associated with the matching path in student lattice is then used as reference state sequence. The approach can be summarized as:

1. Using a shared weak LM we generate lattices for teacher and student models.
2. Teacher lattices are used to find the best word sequence $\mathbf{W}_u^T$.
3. Compute minimum edit distance between $\mathbf{W}_u^T$ and all paths in student lattice. For a path $\pi = (\mathbf{S}, \mathbf{W})$, edit distance is computed between $\mathbf{W}_u^T$ and $\mathbf{W}$.
4. The state label sequence of the path with the least edit distance is used in sMBR training.

3. Experiments

3.1. Data selection

Experiments were performed using anonymized data drawn from Alexa family of devices. ASR confidence scores, described in [25], were a key attribute for unlabeled data selection. For our experiments utterances were selected having an average token confidence between 0.4 to 0.9. This range of confidence scores selects data over a wide distribution. Higher confidence scores correspond to "simpler" utterances with fewer ASR errors and vice versa for lower confidence scores. Apart from token confidence following additional filters were also employed to diversify the information content of selected utterances:

- Utterances marked only with wakeword ("Alexa") intent were filtered out from unlabeled data selection.
- To diversify data across devices, a maximum of 10 utterances were selected from any given device.
- To increase diversity of content in the final selection, we discarded any utterance if similar content had already been selected for 10 times.

Almost 38000 hours of unlabeled and anonymized British dialect data was sampled from production data firehose using this data selection technique.

For CE and CTC training we used an in-house anonymized multidialect dataset composed of four English dialects - American (US), British (UK), Indian (IN) and Australian (AU). This dataset has 45000 hours of labeled data. We also employed data augmentation based on simulated room acoustics [26] to double the size of our initial labeled training data to 90000 hours. A strong supervised sMBR baseline system was trained on natural and simulated 22000 hour subset of British English data. For evaluation we used a test-set containing both clean and noisy far field conditions. The evaluation set consists of 30 hours of clean speech and 10 hours of noisy speech. Decoding was done using a statistical LM combined with a set of domain-specific grammars. We report results as relative Word Error Rate Reduction (WERR) compared to strong supervised baseline system.

3.2. Experimental Details

We now give details of the experimental setup used in this work, followed by a description of various experiments conducted to validate the proposed approach.

Table 1, describes the extracted features and model architecture. All our experiments are based on low frame rate HMM-LSTM hybrid sytem [8, 27]. Signal processing front-end consists of computing Short-time Fourier transform (STFT) at a standard frame rate of 10ms. Log magnitude of the extracted STFT was represented as a 256 dimensional feature vector. Standard mean variance normalization of STFT features was done prior to training. At any given time t, features were stacked with additional two feature vectors from the immediate left to yield a 768 dimensional feature vector.

The target acoustic units were modeled as context dependent (CD) single state triphone units. Choice of CD single state model is motivated by observations in [8, 28] where it was shown that for recurrent architectures CD single state triphone model can equal the performance of a 3-state CD triphone model. The number of output targets determined by phonetic decision tree clustering was 2607 . For CTC training an additional blank symbol was appended to the target list so that the number of CTC targets becomes 2608.

The model architecture used for training is based on time-frequency modeling of speech signal. The STFT features were first passed through a fLSTM network that operates on 768 dimensional feature vector to produce a summary embedding of spectral information at a given time frame. The details of the fLSTM architecture are shared in Table 1. The temporal properties of speech were modeled as recurrence in time using as stacked 5 layer LSTM network. Distributed training was used to train both Teacher and Student models. The networks were initialized with a learning rate of $8e^{-4}$ and Adam optimizer was used as the Stochastic Gradient descent algorithm. Cross-entropy and CTC models were used as seeds for sMBR training. Baseline sMBR reference state labels were generated using force alignment with ground truth text. KD-sMBR techniques discussed in Section 2 were used for sMBR training of unlabeled data. The seed models for sMBR training were initialized with a learning rate of $5e^{-6}$ and a minibatch-size was set to 20480. The non-speech accuracy weight was set to 0.3. For all sMBR experiments we trained over a pool of 16 GPUs using Gradient Threshold Compression method [29].

Feature representation	3 * 256 dimensional [8] Short-Time Fourier Transform
Label representation	2607 Senones (CD Phones) [27]
Student architecture	FLSTM [30] Frequency LSTM : Bidirectional, Window = 48, Hop = 15, Layers = 2, Units = 16. Time LSTM: Unidirectional, Layers = 5, Units = 768
Teacher architecture	FLSTM [30] Frequency LSTM : Bidirectional, Window = 48, Hop = 15, Layers = 2, Units = 16 Time LSTM: Bidirectional, Layers = 5, Units = 1024
Labeled data	45000 hours
Total training data (Labeled + Simulated)	45000*2 = 90000 hours

Table 1: *Experimental architecture and setup*

3.2.1. CE KD-sMBR

We first give detailed analysis of the proposed technique on CE trained models. In Table 2 and Table 3, CE trained models are used as seed to initialize subsequent sMBR stage. Table 2 presents the performance of supervised teacher and student sMBR models with CE used a baseline. We observe the improvements offered by sMBR training stage in the student model. Teacher model being a BiLSTM offers even higher and expected accuracy improvements.

Method	**Clean Speech WERR (%)**	**Noisy Speech WERR(%)**
CE Baseline	0.0	0.0
CE (LSTM) sMBR	12.4	10.8
Teacher (BiLSTM) sMBR	**28.9**	**25.9**

Table 2: *Relative WER (%) reduction for supervised models. Baseline is trained using CE loss on* 90000 *hour (natural and simulated speech) multidialect data. Labeled sMBR training was done on* 22000 *British English subset.*

Table 3 compares the performance of proposed KD-sMBR method against a strong supervised sMBR baseline. The second column shows data, whether labeled, unlabeled or both is used for sMBR training. In the third column we report reference state label sequence estimation method. In the first row for labeled

data reference states were obtained by force aligning ground truth text. In the second row for unlabeled data reference states are obtained from the optimal path in teacher Lattice, i.e., lattice generated from teacher AM and weak LM. We notice some degradation compared to supervised baseline for clean speech while improvements are observed for noisy speech. In the third row lattices are constructed using the student model, teacher AM is then used to estimate the reference state sequence by acoustically rescoring the student model lattices. In comparison to second row, results reveal that first constraining search graph by student AM and then rescoring them by teacher AM gives better results on both clean and noisy conditions compared to labeled baseline. Finally, in forth row we observe that combining both labeled and unlabeled data further improves the model. In this case, the KD-sMBR stage is initialized with labeled sMBR (baseline) model. These results reveal that proposed KD-sMBR training results in substantial improvements over standard supervised models. Moreover, supervised and semisupervised KD-sMBR training have complimentary characteristics and overall system is better than individual ones.

Method	Data	State labels	Clean Speech WERR(%)	Noisy Speech WERR(%)
sMBR Baseline	Labeled	GT	0.0	0.0
KD-sMBR	Unlabeled	TL	-1.8	9.6
KD-sMBR	Unlabeled	SL	4.2	6.3
KD-sMBR	Both	SL	**6.8**	**10.8**

Table 3: *Comparing WER(%) reduction for KD-sMBR CE AM trained on* 38000 *hours of unlabeled speech against labeled sMBR baseline. Baseline is trained on* 22000 *hours of natural and simulated speech. State labels for sMBR training were estimated from either 1.) Ground Truth (GT) 2.) Teacher Lattice (TL) 3.) Student Lattice (SL)*

3.2.2. CTC KD-sMBR

In following experiments we give detailed analysis of applying KD-sMBR for CTC trained models. Table 4 presents the performance of applying sMBR to CTC models for supervised data. Compared to CE training, CTC training minimizes an utterance level loss function. As such, we notice smaller improvement from labeled sMBR training over the CTC baseline. However, even after CTC training significant gains are observed in sMBR stage due to discriminative training over alternative hypotheses. Unsurprisingly, CTC sMBR teacher is better than student sMBR model.

Method	Clean Speech WERR (%)	Noisy Speech WERR(%)
CTC (Baseline)	0.0	0.0
CTC (LSTM) sMBR	6.2	7.2
Teacher (BiLSTM) sMBR	**24.5**	**22.6**

Table 4: *Relative WER (%) reduction for supervised CTC models. Baseline is trained using CTC loss on* 90000 *hour (natural and simulated speech) multidialect data. Labeled sMBR training was done on* 22000 *British English subset.*

Table 5 presents the results of CTC KD-sMBR experiments. The baseline for comparison (first row) is the labeled CTC sMBR model. The second column shows data, whether labeled, unlabeled or both was used for sMBR training. The third column indicates lattice used for reference state label estimation. In the second row, we report the results when reference states are obtained from best scoring path in the teacher Lattice i.e., lattice generated from teacher AM and weak LM. During training we observed a decrease in training objective and this is also reflected in severe degradation in results. This validates our initial hypothesis that since teacher and student CTC models spike at different time frames, as shown in Figure 1, estimating reference state label sequence directly from teacher lattices is sub optimal for unlabeled CTC sMBR training. In third row, we observe that KD-sMBR shows significant improvement for both clean and noisy conditions when student Lattices are used for reference state sequence estimation. In last row, we report results when both labeled and unlabeled data is used for training. In this case unlabeled KD-sMBR training is initialized with labeled sMBR seed model and further improvement is observed for both clean and noisy speech.

Method	Data	State labels	Clean Speech WERR(%)	Noisy Speech WERR(%)
sMBR Baseline	Labeled	GT	0.0	0.0
KD-sMBR	Unlabeled	TL	-888	-577
KD-sMBR	Unlabeled	SL	5.8	5.7
KD-sMBR	Both	SL	**6.2**	**7.7**

Table 5: *Comparing WER(%) reduction for KD-sMBR CTC AM trained on* 38000 *hours of unlabeled speech against* 22000 *labeled CTC sMBR baseline. State labels for sMBR training were estimated from either 1.) Ground Truth (GT) 2.) Teacher Lattice (TL) 3.) Student Lattice (SL)*

LM	Clean Speech WERR(%)	Noisy Speech WERR(%)	RTF ratio
Unigram (KD-sMBR)	0.0	0.0	1.0
5-gram	-3.4	-0.6	5.62

Table 6: *Influence of LM used for reference state label generation.*

In Table 6, we compared the effect of using the teacher hypothesis generated using strong ngram LM against unigram LM. We analyze the performance in terms of compute time and WERR. Compute time is measured in terms of Real Time Factor (RTF) ratio. In these units generating state label sequence with stronger 5-gram LM is more than 5 times as expensive as those generated using unigram LM. Furthermore, the reduction in computational complexity due to KD-sMBR comes without incurring any accuracy trade-off. In fact, our results reveal that KD-sMBR is effective in terms of both computational complexity and WER.

4. Conclusions

Sequence discriminative training, in particular sMBR, is typically the final stage in training of state-of-the-art speech recognition systems. In this work, we investigated the impact of semisupervised learning for improving the sMBR training of hybrid LSTM-HMM acoustic models. We presented an approach based on sequence level knowledge distillation using a teacher student paradigm to drive sMBR training. The proposed KD-sMBR approach was found to be effective in improving both CE trained and CTC trained student models. We established that by leveraging unlabeled data at scale, KD-sMBR approach can outperform standard labeled sMBR trained models. We also show that both labeled and unlabeled sMBR stages have complimentary information and used in conjunction yield better results. An important observation from our experiments is that teacher best obtained through weak LM is more effective reference to guide sMBR training than using strong LM. This has significant implications for scaling up semisupervised sMBR training.

5. References

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

[2] H. Arsikere, A. Sapru, and S. Garimella, "Multi-dialect acoustic modeling using phone mapping and online i-vectors," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2125–2129. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2881

[3] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 4774–4778. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462105

[4] D. Amodei, S. Ananthanarayanan *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, pp. 173–182.

[5] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[6] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. Interspeech 2018*, 2018, pp. 12–16. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1423

[7] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP '09. USA: IEEE Computer Society, 2009, pp. 3761–3764. [Online]. Available: https://doi.org/10.1109/ICASSP.2009.4960445

[8] H. Sak, A. W. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *CoRR*, vol. abs/1507.06947, 2015. [Online]. Available: http://arxiv.org/abs/1507.06947

[9] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Interspeech 2016*, 2016, pp. 2751–2755. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-595

[10] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *in Proc. Eurospeech*, pp. 2725–2728.

[11] J. Z. Ma, S. Matsoukas, O. Kimball, and R. M. Schwartz, "Unsupervised training on large amounts of broadcast news data," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3, pp. III–III, 2006.

[12] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Proc. Interspeech*, 2016.

[13] L. Lu, M. Guo, and S. Renals, "Knowledge distillation for small-footprint highway networks," *CoRR*, vol. abs/1608.00892, 2016. [Online]. Available: http://arxiv.org/abs/1608.00892

[14] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Proc. Interspeech*, 2017.

[15] G. Kurata and K. Audhkhasi, "Guiding CTC Posterior Spike Timings for Improved Posterior Fusion and Knowledge Distillation," in *Interspeech*, 2019, pp. 1616—1620.

[16] R. Takashima, S. Li, and H. Kawai, "An investigation of a knowledge distillation method for ctc acoustic models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5809–5813.

[17] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 250–257, 2018.

[18] N. Kanda, Y. Fujita, and K. Nagamatsu, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level kullback-leibler divergence," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 69–76, 2017.

[19] B. Li, R. Pang, T. Sainath, and Z. Wu, "Semi-supervised training for end-to-end models via weak distillation," in *Proc. ICASSP 2019*, 2019.

[20] S. H. K. Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 6670–6674. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683690

[21] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013.

[22] R. Pang, T. N. Sainath, R. Prabhavalkar, S. Gupta, Y. Wu, S. Zhang, and C.-C. Chiu, "Compression of end-to-end models," in *Proc. Interspeech*, 2018.

[23] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 1317—1327.

[24] G. Kurata and K. Audhkhasi, "Improved knowledge distillation from Bi-directional to Uni-directional LSTM CTC for end-to-end speech recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 411–417.

[25] P. Swarup, R. Maas, S. Garimella, S. H. Mallidi, and B. Hoffmeister, "Improving ASR Confidence Scores for Alexa Using Acoustic and Hypothesis Embeddings," in *Proc. Interspeech 2019*, 2019, pp. 2175–2179. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1241

[26] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[27] G. Pundak and T. N. Sainath, "Lower frame rate neural network acoustic models," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 22–26. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-275

[28] A. Senior, H. Sak, and I. Shafran, "Context dependent phone models for lstm rnn acoustic modelling," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4585–4589.

[29] N. Strom, "Scalable distributed dnn training using commodity gpu cloud computing," in *Proc. Interspeech*, 2015.

[30] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "Lstm time and frequency recurrence for automatic speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 187–191.