EFFICIENT LARGE SCALE SEMI-SUPERVISED LEARNING FOR CTC BASED ACOUSTIC MODELS

Prakhar Swarup, Debmalya Chakrabarty*, Ashtosh Sapru, Hitesh Tulsiani*
Harish Arsikere, Sri Garimella

Amazon Alexa, Bangalore, India

ABSTRACT

Semi-supervised learning (SSL) is an active area of research which aims to utilize unlabeled data to improve the accuracy of speech recognition systems. While the previous studies have established the efficacy of various SSL methods on varying amounts of data, this paper presents largest ASR SSL experiment ever conducted till date where 75K hours of labeled and 1.2 million hours of unlabeled data is used for model training. In addition, the paper introduces couple of novel techniques to facilitate such a large scale experiment: 1) a simple scalable Teacher-Student based SSL method for connectionist temporal classification (CTC) objective and 2) effective data selection mechanisms for leveraging massive amounts of unlabeled data to boost the performance of student models. Further, we apply SSL in all stages of the acoustic model training, including final stage sequence discriminative training. Our experiments indicate encouraging word error rate (WER) gains up to 14% in such a large transcribed data regime due to the SSL training.

Index Terms— Speech Recognition, Semi-Supervised learning, Connectionist Temporal Classification, Teacher-Student

1. INTRODUCTION

Long Short Term Memory (LSTM) RNNs [1] have been shown to outperform feed-forward neural networks [2] and significantly improve the performance of automatic speech recognition (ASR) systems. It has been shown that LSTMs when trained with CTC loss followed by sequence discriminative training can achieve state-of-the-art performance [3, 4]. However, a primary bottleneck of these frameworks is their dependence on large amounts of labeled data. Since collection and transcription of large amounts of speech data is expensive and time-consuming, techniques that additionally leverage unlabeled data for AM training are explored under the SSL framework [5, 6].

In a recent large scale SSL study [7] with 7K hours of labeled data and 1 million hours of randomly selected unlabeled data, it was shown that an AM trained using cross-entropy (CE) on entire data yielded 10-20% relative WER improvement over a baseline model trained only on labeled data. Since randomly selecting data may not be optimal, there have been past studies pertaining to confidence-measure based data selection approaches in the context of small scale SSL [8, 9, 10, 11, 12]. This paper addresses how to effectively select unlabeled data for large scale SSL training with a prime focus on "quality" and "diversity" of unlabeled data. We propose a novel enhanced SSL data selection strategy that: 1) Samples data from various confidence bins based on accuracy improvement due to adding data from each bin and 2) Ensures content diversity by having good coverage across domains and speakers. Our experimental results show that the proposed data selection results in an order of magnitude less SSL data and yet does not degrade ASR accuracy.

Generating labels corresponding to unlabeled data is another crucial aspect of any SSL method. Teacher-Student learning [13] is a dominant SSL paradigm in which a strong offline Teacher model trained on labeled data is used to generate labels for unlabeled data, which are then combined with labeled data. The resultant training set is used for training a student model, which is configured as per the required AM architecture. The Teacher-Student learning is originally proposed for model compression and is also known as Knowledge Distillation (KD). The effectiveness of KD has been well established on several ASR tasks [14, 15, 16, 17].

Existing KD approaches for CTC-SSL training like frame-level [18] or sequence-level KD [19] are based on either window based KL-divergence between teacher and student outputs or generating N-best hypothesis using beam decoding. While such techniques work well with small-scale data (1-2K hours), we believe they may not scale well to massive amounts of unlabeled data due to their space/time complexity. This paper proposes a novel simplified sequence-level KD technique in the context of large scale SSL. The key idea is to first generate a temporary label sequence by concatenating frame-level "argmax" outputs from the Teacher model and then create final label sequence by removing consecutive repetitions and blanks. Our experiments indicate that the proposed simplified sequence-level KD technique significantly

*Equal contribution

reduces the label generation time by eliminating an expensive ASR decoding step and also yields WERs comparable to that with traditional sequence-level KD.

Typical AM training consists of multiple stages, such as CTC/CE followed by sMBR. In most prior work, SSL techniques are independently applied to either the first stage CTC/CE training [7] or the last stage sMBR training [20]. In this work, we apply SSL in all stages of AM training and empirically show that the gains are complimentary. Although it is well-known that increasing the amount of either labeled and/or unlabeled data used in SSL framework yields better results, it remains unknown as to what happens when we push those boundaries. In this paper, we report results of largest ASR SSL experiment ever conducted till date by leveraging 75K hours of labeled and 1.2 million hours of carefully selected unlabeled data. It is interesting to note that our experimental results show a relative WER improvements of up to 14% even in such a large data regime.

The novelty/contributions of this paper are listed below:

1. Propose an effective SSL data selection strategy that selects an order of magnitude less SSL data without degrading the ASR accuracy.

2. Present highly scalable sequence-level KD technique for CTC, which uses frame-level "argmax" outputs from Teacher model for label generation.

3. Demonstrate that applying SSL in all stages of AM training (CTC and sMBR) is complementary and improves ASR accuracy.

4. Conduct largest SSL experiment ever reported by leveraging 75K hours of labeled and 1.2 million hours of unlabeled data to establish efficacy of our solution.

2. DATA SELECTION STRATEGY

Majority of data selection approaches studied in the past have been used for bootstrapping initial seed model with additional unlabeled data [8, 9, 10, 11, 12, 21] in the context of self-learning framework. To the best of our knowledge, data selection in the context of teacher-student based SSL has not been explored extensively.

Utterance-level confidence scores, as described in the baseline setup in [22], are used as one of the attributes for unlabeled data selection. A logistic regression classifier trained on decoder features of an utterance is used to estimate the posterior probability of ASR hypothesis being correct (WER=0). Based on confidence score, unlabeled utterances can be qualitatively classified into "easy" and "difficult". Higher confidence score corresponds to "easy" utterances with fewer ASR errors and vice-versa for lower confidence scores. In this work, we study the effect of confidence on CTC-SSL performance and demonstrate that incorporating SSL information using mid-easy to difficult utterances helps in boosting the performance of ASR system.

Utterance-level domain information, estimated by in-house Natural Language Understanding (NLU) system, is also used for selecting data. This is because ASR accuracies vary across domains (such as Music, Weather and Information etc.) due to varying complexity of utterances across domains as well as varying degree of representation in labeled dataset. We demonstrate the effect of selecting domain-specific SSL data on per-domain ASR accuracies.

Apart from confidence and domain based selection, additional filters are also employed to optimize unlabeled data:

- **No Wakewords**: Wakeword-only ("alexa") utterances are filtered out as they do not add content diversity.
- **Max samples per Content**: We sample a maximum of 50 utterances having identical 1-best recognition result.
- **Max samples per Device**: To increase the number of devices and diversity, we select a maximum of 50 utterances per device. This ensures that frequently used devices do not dominate overall unlabeled data.

3. LABEL GENERATION FOR KNOWLEDGE DISTILLATION IN CTC

One of the tasks of speech recognition involves mapping a frame-level label sequence called 'path' $\boldsymbol{\pi}$ into a label sequence $\boldsymbol{h}$ of length equal to or less than the number of frames. In CTC [3], a path is converted into a label sequence by deletion of repeated as well as blank labels. This conversion is called 'CTC mapping' B, where $\boldsymbol{h} = B(\boldsymbol{\pi})$. Since multiple possible paths can be mapped into a label sequence, conditional probability of label sequence $\boldsymbol{h}$ given input sequence $\boldsymbol{X}$ is defined by Equation (1).

$$P(\boldsymbol{h}|\boldsymbol{X}) = \sum_{\boldsymbol{\pi} \in B^{-1}(\boldsymbol{h})} P(\boldsymbol{\pi}|\boldsymbol{X}) \tag{1}$$

CTC objective is to minimize negative log-likelihood of label sequence given input frames $L_{CTC} = -ln(P(\boldsymbol{h}|\boldsymbol{X}))$.

For SSL study, we need an automatic label generation mechanism for unlabeled data. Based on the model used for label generation, semi-supervised learning methods can be broadly classified into Self-training or Teacher-Student learning. In self-training, a model trained on labeled data uses its own predictions to automatically label the unlabeled examples and use them in training. A constraint in such a framework is the accuracy of generated labels which can get affected by the performance of pre-trained model. For digital assistants like Alexa, the architecture of pre-trained model is restricted since it has to function in an online streaming recognition mode.

In teacher-student based KD, the objective is to train a smaller student model to match the output distribution of a stronger offline teacher model. Since the teacher is not constrained to work in an online scenario, the quality of generated labels

in KD can be improved significantly by using complex architectures for the teacher model. Past studies in CTC have applied KD both at the frame-level [23, 18] and sequence-level [24, 19]. In this work, we focus on sequence-level KD since it was demonstrated in [19] to outperform frame-level KD due to the 'alignment-free' nature of CTC.

In sequence-level KD, student model is trained by minimizing cross entropy loss between probability distributions of label sequences $\boldsymbol{h}$ generated by teacher (P_T) and student (P_S) model as shown in Equation (2).

$$L_{CTC-KD_{seq}} = -\sum_{\boldsymbol{h}} P_T(\boldsymbol{h}|\boldsymbol{X}) ln(P_S(\boldsymbol{h}|\boldsymbol{X})) \quad (2)$$

Since, extracting $P_T(\boldsymbol{h}|\boldsymbol{X})$ for all possible hypotheses is unrealistic, it can be approximated by a set of N-best hypotheses as suggested in [19]. One way to generate such N-best hypotheses is to decode unlabeled utterances using a strong offline teacher acoustic model and a strong language model. However, decoding such large amounts of unlabeled data is a time consuming and expensive effort. In order to simplify label generation, we apply the greedy decoding approximation suggested by Equation (4) in [3] which approximates the best path by concatenating the most probable label at each time step. Overall, this approximation to sequence-level knowledge distillation can be described in following steps:

1. Generate frame-level posterior sequence for each of unlabelled example by running forward pass over strong offline teacher AM.
2. Sample frame level labels π_t^* from frame-level posterior sequence by taking an argmax over teacher posterior vector at each time step t.
3. Student model uses CTC loss for training, which by definition accounts for blanks and repetition of labels. Hence, frame-level representation is compressed by applying CTC mapping operator $B(\boldsymbol{\pi}^*) = \boldsymbol{h}^*$.

$$L_{CTC-KD_{seq}} \approx -ln(P_S(\boldsymbol{h}^*|\boldsymbol{X})) \quad (3)$$

Combining loss for labelled data (D_L) and unlabeled data (D_U), we get the overall student loss function.

$$L = -\sum_{(\boldsymbol{X},\boldsymbol{h})\in D_L} ln(P_S(\boldsymbol{h}|\boldsymbol{X})) - \sum_{\boldsymbol{X}\in D_U} ln(P_S(\boldsymbol{h}^*|\boldsymbol{X})) \quad (4)$$

We measured compute time for generating labels for 1 hour of unlabeled data using each of the above mentioned strategies and quantified it in terms of Real Time Factor (RTF) ratio in Table 1. Table 1 suggests that self-training and full-decoding using a strong teacher and LM are more than 2 times as expensive as proposed argmax based label generation. In the results section, we provide further evidence that the reduction in computational complexity due to proposed strategy comes with minimal WER trade-off.

Label generation strategy	RTF ratio
Self-training	0.4
1-best decoding labels via teacher and LM	0.5
Argmax labels via teacher	0.2

Table 1: RTF comparison across label generation strategies

4. EXPERIMENTS AND RESULTS

Experimental setup used in this study is described in Table 2. The training/evaluation data for building/evaluating acoustic models consist of anonymized labeled dialect-specific English data from in-house data warehouse. For WER computation, we use Language Models (LMs) trained on dialect-specific labeled data sources. We quantify WER gains across baseline and proposed systems via relative WER reduction (WERR %) due to anonymized WER. Baseline WERR % is always reported as 0.0 for the sake of simplicity.

Feature representation	3 * 256 dimensional [25] Short-Time Fourier Transform
Label representation	2608 Senones (Hybrid CTC-HMM) [4]
Student training strategy	Cross entropy(CE) ->CTC
Student architecture	FLSTM [26] Frequency LSTM : Bidirectional, Window = 48, Hop = 15, Layers = 2, Units = 16. Time LSTM: Unidirectional, Layers = 5, Units = 768
Dialects	British English (en-GB) Indian English (en-IN)
Teacher training strategy	Cross entropy ->CTC ->sMBR (Dialect data)
Teacher architecture	FLSTM [26] Frequency LSTM : Bidirectional, Window = 48, Hop = 15, Layers = 2, Units = 16 Time LSTM: Bidirectional, Layers = 5, Units = 1024
Teacher training corpus	75K hours en-US = 45K hours en-GB = 16K hours en-IN = 8K hours en-AU = 6K hours

Table 2: CTC-SSL experimental setup

4.1. Label generation experiment

We present empirical analysis for demonstrating the effectiveness of argmax labels for teacher-student based CTC-SSL. This empirical analysis is presented for en-IN only. Baseline system is trained on 8K hours of en-IN labeled data using strategy mentioned in row 3 of Table 2. In a self-learning framework, the student model itself is used for generating SSL targets. The student model generates 1-best hypothesis using an en-IN specific LM for 47K hours of en-IN unlabeled data. The hypotheses are then converted to senone targets by running a force-alignment step. These senone targets are then used for retraining the baseline student CTC model by interleaving with 8K hours of en-IN labeled data.

Using the same 47K hours unlabeled data, we build two versions of teacher-student based CTC-SSL models. In the first version, a full decoding pass is run using en-IN teacher model and LM to generate 1-best hypothesis, from which senone targets are generated via force alignment. In the second version, we generate argmax labels using the teacher model and technique discussed in Section 3. For both versions, baseline student CTC models are retrained by interleaving unlabeled data with 8K hours of en-IN labeled data.

Table 3 shows that both the teacher-student based CTC-SSL frameworks outperform self-training SSL by yielding additional WERR. Furthermore, rows 3 and 4 suggest that SSL with argmax labels can give WERR gains comparable to fully decoded 1-best hypothesis, while simultaneously attaining framework simplicity. Since argmax label generation doesn't require an LM, it avoids operations like HCLG composition which makes label generation computationally less intensive. For future experiments, we proceed with argmax labels and extend the study to en-GB and en-IN dialects.

Model	WERR(%)
Baseline	0.0
Self-training	9.3
Teacher-Student via 1-best decoding labels	**16.1**
Teacher-Student via argmax labels	**14.6**

Table 3: WERR comparison between Teacher-Student and Self-training based SSL models for en-IN

4.2. Data selection experiments

In following sub-sections, we describe the data-selection experiments that were conducted based on utterance confidence and domain distribution. In both of these experiments, 16K hours and 8K hours of labeled data were used to train baseline CTC models for en-GB and en-IN respectively as per the strategy mentioned in row 3 of Table 2 .

4.2.1. Utterance confidence based sampling

We provide an empirical analysis of selecting unlabeled data based on confidence distribution for CTC-SSL training. The entire unlabeled data is divided into 10 uniformly spaced bins within [0-1000] ([0-1] scaled by 1000) based on confidence values extracted from in-house data warehouse. We sample 5K hours of unlabeled data within each bin for en-GB. However, due to anonymized data sparsity constraints, 8K hours of unlabeled data are sampled within [500-1000] for en-IN. Figure 1 quantifies the effect of confidence based unlabeled data selection from each bin on the performance of CTC-SSL system (in terms of WERR) compared to baseline trained on labeled data only. Note that apart from confidence, other filters discussed in Section 2 (No wakewords, Max samples per content, Max samples per device) are also applied in this study.

Figure 1 clearly suggests that adding unlabeled data from

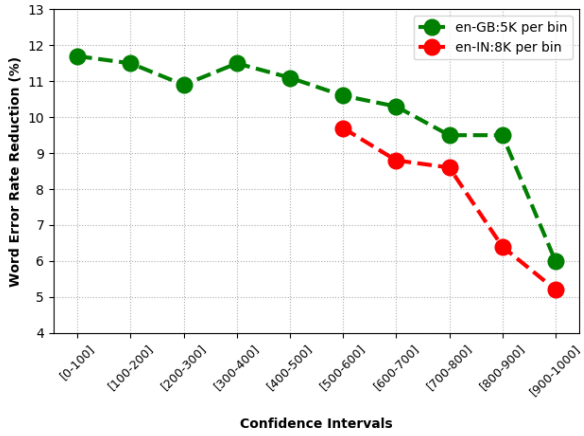

Fig. 1: WERR comparison after adding unlabeled data from utterance confidence bins

low to mid range confidence distribution to CTC training positively impacts the performance of ASR system compared to baseline. As the sampling of unlabeled data moves from low to high confidence, the contribution of SSL data to student performance drops from 11.7% to 6% for en-GB and 9.7% to 5.2% for en-IN. With respect to student model, samples drawn from high confidence bins can be considered as 'easy' and hence adding them to existing labeled data is not beneficial. On the other hand, mid to low confidence bins contain 'difficult' samples which when added to student model training could enable better generalization.

We attempt to devise a strategy to effectively combine data across confidence bins. The approaches explored are:

- **Random sampling (RS)**: Random selection across confidence bins. Filters from Section 2 not applied.
- **Natural distribution (ND)**: Sampling based on natural distribution from in-house data warehouse
- **Uniform distribution (UD)**: Sample equally from all

bins.

- **Weighted sampling (WS)**: Sample based on WERRs in Figure 1

The results listed in Table 4 show that 40K hours of unlabeled data combined using WS and UD methods yield better results than that of ND. Interestingly, these unlabeled data selection methods with order of magnitude less data yield comparable results to 250K hours of randomly sampled data which reinforces the importance of data selection for SSL.

Model	en-GB WERR (%)	en-IN WERR(%)
Baseline	0.0	0.0
SSL with 250k [RS]	17.2	17.6
SSL with 40k [ND]	15.1	14.1
SSL with 40k [UD]	17.1	**17.3**
SSL with 40k [WS]	**17.4**	16.3

Table 4: WERR comparison between different confidence bin based combination methods

4.2.2. Utterance domain based sampling

We present empirical analysis of the effect of domain based sampling of unlabeled data in boosting ASR performance. Based on NLU domain extracted from in-house data warehouse, we sample 10K hours and 8K hours from 5 individual domains [D1$\cdots$D5] in en-GB and en-IN respectively and build CTC-SSL model by interleaving labeled data with argmax labels from individual domain's unlabeled data. Apart from domain, other filters discussed in Section 2 (No wakewords, Max samples per content, Max samples per device) are also applied in this study.Table 5 validates the hypothesis that individual domain sampling improves the performance of corresponding domain without significant degradation across other domains, as supported by the diagonal behavior in en-GB and en-IN WERR matrices.

Small cross domain degradation is observed e.g. [D5 model-D2 test] in en-GB and [D5 model-D3 test] in en-IN. We speculate that boosting a particular domain-specific data via SSL may result in data imbalance in training set. To tackle this issue, data across all the selected domains was combined (8K hours * 5) such that each domain gets enough representation during CTC-SSL training. This 40K hours of domain data ingested into training provides the greatest WERRs across all domains for en-IN as can be observed from comparing last row in Table 5 against the other rows. However, for en-GB the per-domain WERRs from combined domain data are lesser than those obtained from individual domain based sampling. This is because for en-GB, individual domain based experiments had a better representation per-domain (10K hours) as opposed to the combined domain experiment (8K hours * 5).

Model	en-GB WERR (%)					en-IN WERR (%)				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Baseline	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D1	**16.8**	5.0	-1.5	4.5	10.6	**12.1**	6.5	2.2	11.6	0.5
D2	5.5	**12.2**	5.2	7.3	12.8	8.2	**10.0**	3.7	7.7	6.6
D3	1.6	0.6	**21.2**	5.7	7.1	5.6	4.7	**12.5**	7.7	4.5
D4	6.8	8.0	14.7	**11.1**	8.6	8.1	5.8	4.3	**15.5**	2.0
D5	10.1	-1.7	-0.8	5.2	**21.9**	7.2	1.7	-4.1	5.7	**8.5**
Combined Domains	12.4	10.3	20.9	7.7	14.7	16.4	12.9	15.7	22.1	13.28

Table 5: WERR comparison across CTC-SSL models based on individual domain based sampling

4.3. Multi-stage AM training using SSL

We present the effect of our data selection based SSL strategies on a large scale multi-stage CTC-sMBR based AM setup and demonstrate that adding SSL information across each of these stages provide complimentary information and lead to additive gains compared to a strong baseline system.

4.3.1. Baseline AM

For building a strong baseline system, we firstly build a multi-dialect seed CTC model by pooling labeled data across multiple dialects (en-US, en-GB, en-IN and en-AU) which amounts to 75K hours of data (Row 8 in Table 2) as suggested in [27]. Data augmentation via simulated reverb addition is used to double the amount of labeled data to 150K hours. Using multi-dialect CTC model as the seed, we perform a two-step fine-tuning training using dialect specific labeled data. In the first step, a few additional epochs of CTC fine-tuning is run over labeled data and followed by few epochs of sMBR (Sequence-level Minimum Bayes Risk [28]) in the second step. Such a training strategy establishes a strong baseline for individual dialects as suggested by results shown in Table 6 where final sMBR training led to 8.9% and 8.2% WERR over pooled multi-dialect CTC model for en-GB and en-IN respectively. This final sMBR model is treated as baseline system for individual dialects and compared against our proposed SSL based AM framework.

4.3.2. Multi-stage SSL AM

We add **carefully selected** unlabeled data across multiple dialects (en-GB=40K hours, en-IN=40K hours, en-AU=20K hours) to replicate the pooled training setup as done in case of baseline. "Carefully selected" refers to **UD** based data selection as established via our experimental findings in Section 4.2. With this sampling approach, we obtained a 6% WERR improvement at pooled training stage compared to natural distribution based sampling, hence establishing the efficacy of such a data selection framework in a large scale SSL data regime. A brief summary of data selection criteria used in this experimental setup is mentioned below:

- Utterance confidence: Uniform sampling in (0, 800)

- No wakewords
- Max utterances per content: 50
- Max utterances per speaker-domain: 50

For SSL training, argmax labels are generated for unlabeled data using a strong offline teacher and then interleaved with labeled targets. Pooled model is trained with 250K hours of data (75K*2 labeled + 100K SSL), which is further used as a seed for fine-tuning both en-GB and en-IN dialects. The pooled CTC-SSL model is fine-tuned on individual dialects by running additional few runs of epochs of CTC training on dialect specific labeled (en-GB=16K hours and en-IN=8K hours) and SSL data(40K hours for both en-GB and en-IN). In the final sequence discriminative training stage, a few additional epochs of sMBR training is performed over respective fine-tuned models using labeled data only for en-GB and en-IN. From the results reported in Table 6, comparison between row 4 and 5 clearly suggests that even against a strong baseline system, our data selection based SSL strategy at CTC stage imparts significant WERR for both en-GB and en-IN in the range of $4-7\%$, hence demonstrating that the quality and diversity of SSL data adds complimentary information even in large scale data regime.

Model	Training stage	en-GB WERR (%)	en-IN WERR (%)
Baseline AM (Only labeled)	Pooled	0.0	0.0
	Fine-tuning	1.2	3.7
	sMBR	8.9	8.2
SSL AM (Labeled + SSL)	Baseline	0.0	0.0
	CTC-SSL + sMBR	3.8	6.9
	CTC + sMBR-SSL	2.0	3.0
	CTC-SSL + sMBR-SSL	5.7	8.8

Table 6: WERR comparison between baseline and SSL based multi-stage AM

In a recent work [20], it was demonstrated that teacher-student based KD-sMBR approach helps in improving both CE and CTC trained student models and can outperform standard labeled sMBR trained models. KD-sMBR is a distillation approach for sequence discriminative training where reference state sequence for unlabeled data are estimated using a strong Teacher model. To demonstrate the complementarity of unlabeled data at CTC and sMBR stage of training, we build two versions for comparison. In the first version, we use a transcribed only CTC model (pooled and fine-tuned) as the seed for sMBR training and incorporate unlabeled data only via KD-sMBR. In the second version, we incorporate SSL information at all stages of AM training: pooled, fine-tuning and sMBR. We employed two pass sMBR strategy in which the first pass is regular sMBR training on labeled data and second pass is KD-sMBR training using unlabeled data.
Results from both versions are reported in rows 6 and 7 of Table 6. Row 6 clearly shows that incorporating SSL information only at sMBR stage of multi-stage AM via teacher-student KD-sMBR provides $2-3\%$ WERR over baseline systems for both en-GB and en-IN. By further adding SSL at all training stages, the final system provides an overall WERR of $6-9\%$ as reported in row 7 which shows the complimentary nature of SSL at CTC and sMBR stages.

4.4. Large Scale SSL based AM

We increased the amount of SSL data for individual dialects such that en-GB amounts to 880K and en-IN amounts to 330K hours of carefully selected unlabeled data. The primary motivation behind this study is to understand how far the boundary can be pushed using carefully selected data for achieving maximum WERR. Following a similar SSL strategy as explained in last section, we build a large scale SSL based multi-stage AM and compare it against baseline systems built using only labeled data and SSL AM built using 100K hours of SSL data, both of which are described in last section. All the systems are being compared at final sMBR stage. From comparison reported in Table 7, we see that large scale SSL based AM brings additional WERRs in range of $2-6\%$ compared to other two reported systems, hence establishing the fact that large scale carefully selected SSL data can still add to the diversity of an already existing well selected data regime.

Model	en-GB WERR (%)	en-IN WERR(%)
Baseline	0.0	0.0
SSL with 100 KHrs	5.7	8.8
SSL with 1.2 Mhrs	7.5	14.4

Table 7: WERR comparison across different data scales

5. CONCLUSIONS

This paper presents an empirical study on large-scale semi-supervised learning for CTC acoustic models where a strong offline teacher model is used to generate labels for unlabeled data. The unlabeled data is selected based on confidence and domain distribution as well as speaker and content variability. Experimental results on two different dialects reinforce the efficacy of teacher generated argmax labels and the importance of intelligent data selection methods. It is observed that low to mid confidence ranges are important for reducing WER and that domain-specific unlabeled data has a strong impact on corresponding WER with little cross-domain impact. In a large scale, multi-stage AM training setup, we were able to successfully apply these data selection strategies to achieve a WERR of $8-14\%$ over strong baseline systems for 2 English dialects. Future work in this direction would be to devise a strategy to leverage both confidence as well as domain diversity in a combined data sampling strategy for SSL. Another important direction would be to explore the effect of N-best scalable sampling strategies using different teacher distributions for boosting the student performance on a large scale AM setup.

6. REFERENCES

[1] Alex Graves and Jürgen Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, vol. 18, no. 5, pp. 602–610, 2005.

[2] Hasim Sak, Andrew W. Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4280–4284.

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

[4] Andrew Senior, H. Sak, F. de Chaumont Quitry, T. Sainath, and K. Rao, "Acoustic modelling with CD-CTC-SMBR LSTM RNNs," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 604–609.

[5] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4844–4848.

[6] Yan Huang, Yongqiang Wang, and Yifan Gong, "Semi-supervised training in deep learning acoustic model," in *INTERPSEECH*, 2016, pp. 133615–133627.

[7] S. H. Krishnan 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, pp. 6670–6674.

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

[9] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6704–6708.

[10] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433 – 444, 2010.

[11] T. Tsutaoka and K. Shinoda, "Acoustic model training using committee-based active and semi-supervised learning for speech recognition," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2012, pp. 1–4.

[12] F. d. C. Quitry, A. Oines, P. Moreno, and E. Weinstein, "High quality agreement-based semi-supervised training data for acoustic modeling," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 592–596.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[14] Suyoun Kim, Michael L. Seltzer, Jinyu Li, and Rui Zhao, "Improved training for online end-to-end speech recognition systems," in *INTERSPEECH*, 2018, pp. 2913–2917.

[15] Yevgen Chebotar and Austin Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *INTERSPEECH*, 2016, pp. 3439–3443.

[16] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey, "Student-teacher network learning with enhanced features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5275–5279.

[17] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *INTERSPEECH*, 2017, pp. 3697–3701.

[18] 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.

[19] 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.

[20] Ashtosh Sapru and Sri Garimella, "Leveraging unlabeled speech for sequence discriminative training of acoustic models," in *INTERSPEECH*, 2020.

[21] Langzhou Chen and Volker Leutnant, "Acoustic Model Bootstrapping Using Semi-Supervised Learning," in *INTERSPEECH*, 2019, pp. 3198 – 3202.

[22] Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister, "Improving ASR confidence scores for alexa using acoustic and hypothesis embeddings," in *INTERSPEECH*, 2019, pp. 2175–2179.

[23] Gakuto Kurata and Kartik Audhkhasi, "Guiding CTC Posterior Spike Timings for Improved Posterior Fusion and Knowledge Distillation," in *INTERSPEECH*, 2019, pp. 1616 –1620.

[24] Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu, "Knowledge distillation for sequence model," in *INTERSPEECH*, 2018, pp. 3703–3707.

[25] Hasim Sak, Andrew W. Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *INTERSPEECH*, 2015, pp. 1468–1472.

[26] Tara Sainath and Bo Li, "Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks," in *INTERSPEECH*, 2016, pp. 813–817.

[27] Harish Arsikere, Ashtosh Sapru, and Sri Venkata Surya Garimella, "Multi-dialect acoustic modeling using phone mapping and online i-vectors," in *INTERSPEECH*, 2019, pp. 2125–2129.

[28] Brian Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3761–3764, 2009.