White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/209220/

Version: Accepted Version

Proceedings Paper:
Ahmad, R., Farooq, M.U. and Hain, T. (2024) Progressive unsupervised domain adaptation for ASR using ensemble models and multi-stage training. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024), 14-19 Apr 2024, Seoul, Korea. Institute of Electrical and Electronics Engineers (IEEE), pp. 11466-11470. ISBN: 979-8-3503-4485-1. ISSN: 0736-7791. EISSN: 2379-190X.

https://doi.org/10.1109/ICASSP48485.2024.10448438

© 2024 The Author(s). Except as otherwise noted, this author-accepted version of a conference paper published in International Conference on Acoustics, Speech, and Signal Processing (ICASSP) is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

PROGRESSIVE UNSUPERVISED DOMAIN ADAPTATION FOR ASR USING ENSEMBLE MODELS AND MULTI-STAGE TRAINING

Rehan Ahmad, Muhammad Umar Farooq, Thomas Hain

Speech and Hearing Group, The University of Sheffield, UK

ABSTRACT

In Automatic Speech Recognition (ASR), teacher-student (T/S) training has shown to perform well for domain adaptation with small amount of training data. However, adaption without ground-truth labels is still challenging. A previous study has shown the effectiveness of using ensemble teacher models in T/S training for unsupervised domain adaptation (UDA) but its performance still lags behind compared to the model trained on in-domain data. This paper proposes a method to yield better UDA by training multi-stage students with ensemble teacher models. Initially, multiple teacher models are trained on labelled data from read and meeting domains. These teachers are used to train a student model on unlabelled out-of-domain telephone speech data. To improve the adaptation, subsequent student models are trained sequentially considering previously trained model as their teacher. Experiments are conducted with three teachers trained on AMI, WSJ and LibriSpeech and three stages of students on SwitchBoard data. Results shown on eval00 test set show significant WER improvement with multi-stage training with an absolute gain of 9.8%, 7.7% and 3.3% at each stage.

Index Terms— Automatic Speech Recognition, domain adaptation, pseudo-labeling, pre-training, self-supervised learning

1. INTRODUCTION

The performance of automatic speech recognition (ASR) has been improved dramatically with advanced end-to-end neural network architectures [1]. However, it is shown that these models does not perform well [2] on cross-domain data due to mismatch between training and test data domains. Domain adaptation [3] tries to solve this problem by transferring the model trained on source data to the target domain data. Adaptation is achieved via supervised [4] or unsupervised [5] methods depending on the availability of labelled target data. Domain adaptation method can be straightforward with the presence of labelled target data i.e. fine-tuning the source models on the target data [4, 6]. However, with unlabelled target data unsupervised domain adaptation (UDA) requires specific modelling [7, 5] approaches to efficiently learn the new domain.

Teacher-student (T/S) [8] training is a common approach to transfer knowledge of one model to another. It has been widely used for model compression [9, 10, 11], domain adaptation [3, 12, 13, 14, 15] and generalisation [16]. Ensemble teacher models in T/S training have shown great benefit for acoustic model training [17, 18, 19]. Having multiple teachers, a student can adapt better [20] leveraging diverse teacher outputs. The output of multiple teacher models can be effectively combined to train the student model. One of the simplest approach of combining teacher outputs is taking the average of the posterior distributions from all the teachers [9, 17]. However, previous work [18, 20] has shown that taking the average of posterior distributions from all the teacher models and using them in training a student ASR model does not lead to any improvements in error rate. In [18], authors have tested many selections methods including Top-1, Top-K, averaging and weighted-average with Top-1 being their best approach. Similarly, in [20] we have proposed Top-1 based unsupervised selection method for UDA to select the best posteriors from ensemble of teachers. With the proposed selection method student model is only trained on posteriors with lower error. Hence, the student model performed significantly better compared to all the teachers. However, the performance of the student model still lags behind compared to the state-of-the-art in-domain model.

Recently, iterative [5, 21] and multi-stage [22] training methods have been proposed for acoustic model training. In [22], Rathod et al. has showed that the multi-stage training can be used in T/S method to gradually compress the model without losing significant performance. Similarly, iterative pseudo-labelling [21] and self-iteration methods [5] have shown improvements in semi-supervised and unsupervised data selection for ASR. Inspired by improvements shown in these methods, multi-stage training is exploited in this work to progressively improve domain adaptation.

This paper improves previously proposed UDA in ensemble T/S training method for ASR [20] by integrating multi-stage training. In our previous work [20], multiple teacher models were trained on the source data. These teachers were used to generate soft-labels (posteriors) for the target data during inference. From multiple teachers, soft-labels were selected by unsupervised selection method to train a student model. The results showed significant WER improvement on target data by the student model compared to the teachers. To improve the adaptation, several changes are made in this work which is explained as follows. Since beam-search [23] has been proven to perform better than greedy decoding, therefore in this work soft-labels generate by teacher models are replaced by hard-labels by decoding through beam-search with out-of-domain (OOD) LM. The student model is now trained on the hard-labels. Moreover, as discussed in the previous work that the student performed better than the teachers, therefore in this work the trained student is used as a teacher to improved previously generated pseudo-labels. The new pseudo-labels are then used to train another student model. The process of generating the pseudo-labels by previous student and using them to train next student is repeated to progressively improve the adaptation. The experiments are conducted with three source (labelled) data from meeting and read speech and a target (unlabelled) data from conversational telephone speech. The results show that the student models trained up to three stages show significant WER improvement at each stage with an absolute difference of 9.8%, 7.7% and 3.3%. To study some insights, a comparative analysis of fine-tuning wav2vec-base and wav2vec-large models is conducted at each training stage. Secondly, WER of pseudo-labels generated by teacher and student models is also observed at each training stage.

2. ENSEMBLE AND MULTI-STAGE MODELS

The proposed ensemble teacher-student training paradigm assumes the existence of N teacher models trained independently on different labelled datasets. Each model is based on a pre-trained wav2vec2.0 [24], connected with feed forward network (FFN) layers. Input to the wav2vec2.0 models is a raw waveform represented by X and outputs are the feature representations $\mathbf{c_1}, ..., \mathbf{c_T}$ for T time-steps. These features are fed to the FFN layers to produce the output posterior distributions $\mathbf{h_1}, ..., \mathbf{h_T}$. Each $\mathbf{h_i} \in \mathbb{R}^z$ with a grapheme vocabulary represented by $G = \{g_1, g_2, ...g_z\}$. All the teacher and student models are fine-tuned using CTC [25] loss defined as:

$$\mathcal{L}_{CTC} = -\sum_{B} \log p(\mathbf{y}|\mathbf{h}) \tag{1}$$

where $\mathbf{y}$ represents the grapheme and B is the whole training dataset. During fine-tuning, teacher models use ground-truth labels while students use pseudo-labels generated by the teacher(s).

Outputs of the ensemble teachers are used to train a single student model which is further extended into multi-stage models. In multi-stage training, M student models are trained successively using pseudo-labels generated by the last stage model. The value of M depends on the progressive improvement at each stage which is kept increasing until no improvement is observed in the WER of a subsequent student. Each stage of the student is represented by Sk, where $k \in \{1, 2, ..., M\}$. During inference, each teacher model generates output posterior distributions for the same utterance, they are selected via unsupervised sampling method to train the student. The following section describes the selection strategy in detail. Figure 1 shows the modelling approach comprising of ensemble teacher models and multi-stage students.

2.1. Selection strategy

An unsupervised selection strategy is employed for the ensemble teachers to select the best teacher's output. The selection is made on the basis of average posterior values from each teacher model. In [20], we discussed many selection strategies to select the best output from ensemble of teachers including 'averaging', 'frame-wise' and 'Top-1'. However, the Top-1 strategy has been proved to outperform all other techniques and is employed in this work.

For each input utterance, N teacher models generate output posterior distributions for T time-steps represented as $\mathbf{h_1^1}...\mathbf{h_T^N}$. At each time-step i maximum posterior value for teacher n is selected as:

$$p_i^n = \max_z \mathbf{h_i^n} \tag{2}$$

where z is the vocabulary size.

An average is taken across the time-steps to get an averaged posterior value for each teacher as:

$$q^n = \frac{1}{T} \sum_{i=1}^{T} p_i^n \tag{3}$$

This average value q^n is the score of each teacher n and the teacher with maximum score is selected as:

$$b = \arg\max_n q^n \tag{4}$$

where b is the selected teacher number. The output posterior distributions of teacher b is used to train the student model. With this selection strategy output posterior distribution for each utterance is dynamically selected from one of the teacher models on the basis of their scores.

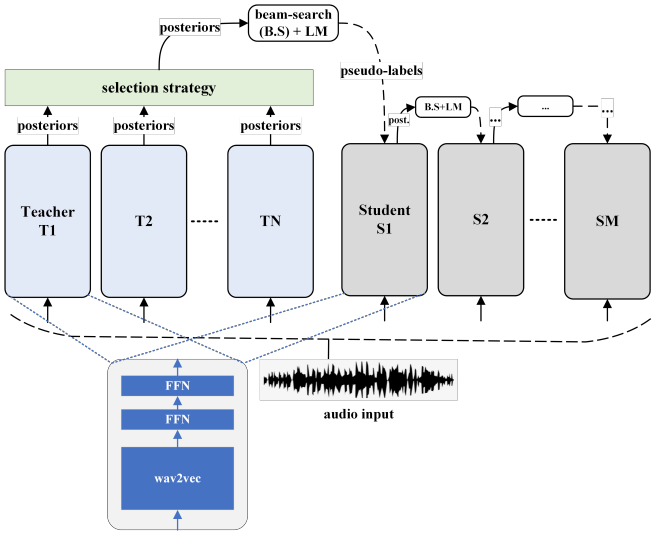

Fig. 1. An ensemble of teacher models with multi-stage students. Each teacher and student model is based on wav2vec with Feed Forward Network (FFN) layers. Each model produces posterior distributions during inference and takes the pseudo-labels during fine-tuning.

2.2. Training method

Figure 1 shows the ensemble teacher models with multi-stage students. N teacher models are trained independently with N source domain labelled datasets. To train a student model, an unlabelled out-of-domain (OOD) target data is used. Since the target data is unlabelled, therefore student is trained using the labels generate by the teacher models.

To train the $S1$ student, all the teachers run inference on the target data and generate output posterior distributions. The posteriors of the best teacher are selected by the selection strategy discussed in the Section 2.1. Selected posteriors are finally converted into pseudo-labels using beam-search[1] integrated with an out-of-domain LM as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \log p_{AM}(\mathbf{y}|\mathbf{x}) + \alpha \log p_{LM}(\mathbf{y}) + \beta|\mathbf{y}| \tag{5}$$

where p_{AM} and p_{LM} represent acoustic and language model probabilities. α and β are hyper-parameters optimised on validation set.

In training the $S1$ model, the supervision is provided by ensemble teacher models. As the $S1$ model training benefited from diverse teacher outputs, therefore it outperforms all the teacher models. This has already been shown in [20]. Once $S1$ is trained, it is used as a teacher to update previously generated pseudo-labels. The updated pseudo-labels are then used to train next stage student $S2$. This process is repeated for M number of stages and training process is stopped when no improvement is noticed in a subsequent student model. The idea behind multi-stage training is that if a student performs better than its teacher then use this student as a teacher to update the previous pseudo-labels. Then, train a next student with the updated labels.

[1] https://github.com/kensho-technologies/pyctcdecode

3. EXPERIMENTS

3.1. Data

Four corpora from different domains have been used to experiment with ensemble multi-stage teacher-student training. These include AMI [26], WSJ [27] (LDC93S6A, LDC94S13A), LibriSpeech (LS) [28] and SwitchBoard (SWBD) [29]. From LibriSpeech, a smaller subset of 360h is taken which is referred as LS360. Durations of AMI, WSJ and SWBD is 100, 272 and 300 hours respectively. SwitchBoard audio data is upsampled to 16kHz for wav2vec2.0 model requirement. In terms of domains, AMI is meeting speech recorded in a meeting room, WSJ and LS360 are read speech data recorded in the studio and SWBD is conversational telephone speech. Among these corpora AMI, WSJ and LS360 are used as source labelled data to train teacher models and SWBD as an unlabelled taget data for the student. Student models are evaluated on the eval00 test set. This test set combines CallHome (LDC97S42) and SwitchBoard test sets represented as CH and SB.

3.2. Experimental setup

In the experiments, three teacher models are considered, i.e $N = 3$, and the student models are trained up to three stages $M = 3$. Teacher models are based on wav2vec-large pre-trained on LV-60K and fine-tuned independently on LS360, WSJ and AMI. For the student model, either wav2vec-base pre-trained with LS960 or wav2vec-large pre-trained with LV-60K is considered for different set of experiments. The student is fine-tuned on SWDB pseudo-labels whose domain is completely mismatched with the teachers. To acquire pseudo-labels and evaluate each student model, a 3-gram language model is trained by combining the transcripts of AMI, WSJ and LS360. As LM is trained on the teachers data, it is also an out-of-domain model for the student.

The first experiment investigates the advantage of training the student model using hard-labels generated by beam-search with OOD LM. In this experiment, ensemble teacher models are used to train the $S1$ student model. The WER of $S1$ model is compared with S_{KL} and S_{Or}. S_{KL} is the first-stage student trained on the soft pseudo-labels using KL loss. S_{Or} is the first-stage student model trained with the oracle selection of the teachers. Oracle selection is made by selecting each posterior distribution which has lowest WER on each training utterance.

In the second experiment, multi-stage training is conducted where $S1$ model is used as a teacher to generate the pseudo-labels of the training data and update the previously generated pseudo-labels. The $S2$ student is then trained using new labels. Similarly, subsequent $S3$ student is trained after $S2$ in a similar manner. All the models are evaluated on eval00, CH and SB test sets with and without using 3-gram language model. In the above experiments, all the models are based on wav2vec-large for the sake of fair comparison. Table 1 shows the WER of all the teachers and students.

The third experiment provides a comparison between wav2vec-base and wav2vec-large self-supervised models which are fine-tuned on the noisy pseudo-labels at each stage of the student model. This experiment aims to analyse if large models with higher capacity in the number of parameters and pre-training data provide any gain compared to the smaller ones when trained with noisy labels. The wav2vec-base model is pre-trained with LS960 and has 95M parameters and wav2vec-large is pre-trained with LV-60K and has 317M parameters. These results are presented in Table 2.

4. RESULTS AND DISCUSSIONS

Table 1 shows the results of the first two experiments. We first report the WER for the three teacher models which are represented by the name of the dataset they were trained on i.e. AMI, LS360 and WSJ. These teachers provide state-of-the art results on their in-domain test sets [20]. SWBD baseline results are also presented to show the in-domain results on eval00. For SWBD baseline, a wav2vec model is fine-tuned using labelled SWBD training data. The in-domain results are shown in shaded cells. All the results in this table are presented with and without an OOD LM. Same LM is used to in beam-search to produce pseudo-labels for each student. Among three teacher models it can be seen that the lowest WER on eval00, CH, SB is provided by LS360 model i.e. 41.8%, 48.6% and 36.7% respectively, which is further improved by OOD LM upto 3.7% absolute. Overall, for eval00 the LS360 model outperform AMI and WSJ by 5.6% and 22.4% reduction in WER.

As the $S1$ model is trained by selecting the best teacher's output using selection strategy, a significant improvement is noticed compared to all the teachers with WER of 32%. For $S1$, the absolute WER improvement over the best teacher (LS360) is 9.8%, 10.8%, and 8.9% for eval00, CH and SB evaluation sets. Compared to the S_{KL} model, which uses soft posteriors and KL loss in training first stage student, the $S1$ model outperforms by 4.2% WER for eval00. This shows the significance of first proposed improvement in the training method by using beam-search with LM to get better training labels for the student model. The selection strategy is an unsupervised method which means there is a chance to select wrong teacher for each utterance therefore an oracle student model S_{Or} is trained to compare with $S1$. The WER of the S_{Or} is 4.6% better than the $S1$, which shows there is a room to improve the selection method however the main challenge is an unsupervised way of selecting best teacher for each utterance. OOD LM lead to further improvement in the WER of the $S1$ from 32.0% to 26.2% for eval00 i.e. 5.8% absolute gain. Similar, results are shown by CH and SB evaluation sets.

The $S1$ model outperforming all the teachers is used to update the pseudo-labels. The $S2$ model is then trained with new pseudo-labels. From table 1, it can be seen that the $S2$ outperforming the $S1$ model with WER of 24.3% (w/o LM) and 21.3% (w/ OOD LM) for eval00. The absolute gain by the $S2$ is 7.7% (w/o LM) and 4.9% (w/ OOD LM). Similarly, the $S3$ model trained on pseudo-labels provided by the $S2$ outperforming it with WER of 21% (w/o LM) and 19.6% (w/ OOD LM) with gain of 3.3% (w/o LM) and 1.7% (w/ OOD LM). Comparing the individual gain at each stage of the student models it is observed that the highest improvement is noticed at first stage and then it gradually decreases as the performance of the model gets closer to the in-domain baseline. In brief, the WER improvement on eval00 at each stage compared to their teachers is 9.8%, 7.7% and 3.3%. WER reduction from LS360 to $S3$ is 20.8% and $S1$ to $S3$ is 11%. These results show the significance of the proposed methods in improving the unsupervised domain adaptation. Compared to our previous method [20], WER of the multi-stage are much closer to the SWBD baseline model.

Table 2 presents the results of the third experiment where the wav2vec-base (w2v-B) and the wav2vec-large (w2v-L) models are trained at each stage and their WER are compared. For the $S1$ model in table 2, the results show that the w2v-L model do not show improvements on eval00 and SB test sets and minimal improvement on CH. On eval00, WER of the w2v-B is 0.5% better than the w2v-L w/o LM and 0.2% w/ LM. Similar difference is noticed for SB test set. However, for CH set the w2v-L is 0.4% better than the w2v-B. At the first stage training of the student model the results show

Table 1. WER(%) evaluation of teachers and multi-stage students on SWBD test sets with and without (OOD) LM. The shaded boxes show in-domain results of SWBD and rest are out-of-domain. S_{Or} is the first stage student trained with oracle selection of posteriors and S_{KL} is the first stage student trained on soft-posteriors (acquired through selection strategy) with KL loss. All the models are based on wav2vec-large.

	Teacher models			**Baseline (in-domain)**	**Student models**				
Test sets	**AMI**	**LS360**	**WSJ**	**SWBD**	**SWBD**				
					S_{Or}	S_{KL}	$S1$	$S2$	$S3$
	w/o LM								
eval00	47.4	41.8	64.2	11.9	27.4	36.2	32.0	24.3	**21.0**
CH	52.0	46.8	71.6	15.3	31.6	40.3	36.0	28.1	**24.5**
SB	42.5	36.7	56.5	8.4	23.1	31.8	27.8	20.3	**17.4**
	w/ LM (OOD)								
eval00	44.3	38.2	61.8	10.1	22.4	28.5	26.2	21.3	**19.6**
CH	49.0	43.2	69.7	12.8	26.1	33.3	30.2	25.0	**23.1**
SB	39.3	33.0	53.5	7.3	18.5	23.6	22.0	17.4	**16.0**

Table 2. Performance comparison of wav2vec-large (w2v-L) and wav2vec-base (w2v-B) models at each stage.

	$S1$		$S2$		$S3$	
Test sets	**w2v-B**	**w2v-L**	**w2v-B**	**w2v-L**	**w2v-B**	**w2v-L**
	w/o LM					
eval00	**31.5**	32.0	24.7	**24.3**	22.9	**21.0**
CH	36.4	**36.0**	29.2	**28.1**	27.6	**24.5**
SB	**26.3**	27.8	**20.1**	20.3	17.9	**17.4**
	w/ LM (OOD)					
eval00	**26.0**	26.2	22.4	**21.3**	20.5	**19.6**
CH	30.6	**30.2**	26.7	**25.0**	24.8	**23.1**
SB	**21.2**	22.0	17.9	**17.4**	16.0	**16.0**

Table 3. WER of the teacher and student models on SWBD training data computed during inference. For the 'Teachers' model, WER is computed after selection strategy.

Model	**WER of Training data**	
	w/o LM	w/ LM (OOD)
Teachers	41.9	38.8
$S1$	32.4	26.8
$S2$	25.5	22.9
$S3$	22.9	22.0

that there is no benefit of using the w2v-L model over the w2v-B model. For the $S2$, some improvements are noticed for the w2v-L compared to the w2v-B with 0.4% (w/o LM) and 0.9% (w/ OOD LM) gain in WER on eval00 and 1.1% (w/o LM) and 1.7% (w/ OOD LM) on CH. This difference is further improved by the $S3$ model with 1.9%, 3.3% on eval00 and CH. The results show that as the subsequent model gets better compared to its teacher and provide better pseudo-labels for the next stage, the w2v-L model starts showing improvement over the w2v-B model. This suggests that the large model with more pre-training data has an advantage over smaller one when trained with less noisy labels. Further details of the WER of the pseudo-labels at each training stage is discussed in the following section.

4.1. Analysis of selection strategy and pseudo-labels

This section aims to provide insights into the the performance of selection strategy and WER of pseudo-labels at teachers and student stages. These analysis are done for the SWBD training data during inference. With the help of oracle selection it is computed that almost 75.5% utterances are correctly selected by selection method which lead to the 32% WER for $S1$ compared to 41.8% of LS360. Student model trained with oracle selection (100% correct) S_{Or} is able to achieve WER of 27.4% as shown in table 1. This indicates that improvement in the selection method would eventually lead to better first stage student and it would be further propagated to the subsequent students.

Table 3 shows the WER of teacher and student models on SWBD training data. Results in this table may help to understand how much pseudo-labels are improving at each training stage leading to better subsequent student model. The results in this table are presented with and without using beam-search and OOD LM. As discussed earlier, pseudo-labels only acquired through beam-search and OOD LM are used in training the next stage model. The results w/o LM are helpful to understand the contribution of OOD LM in improving the pseudo-labels. In table 3, WER of 'Teachers' is computed after the selection strategy. The WER of teacher models after selection is 41.93% which is improved to 38.8% using OOD LM. This means that the $S1$ is trained with pseudo-labels having WER of 38.8%. The WER of the $S1$ model on the SWBD training set is 32.4% (w/o LM) and 26.8% (w/ OOD LM). It can be seen that pseudo-labels provided by $S1$ are significantly improved compared to the teachers i.e. 41.9% (w/o LM) to 32.4% (w/o LM) with 9.5% absolute gain. Using OOD LM the difference between 'Teachers' and $S1$ is increased to 12% with WER of 26.8% (w/ OOD LM). At this stage, there is a significant improvement in the error of pseudo-labels. Similarly, at the $S2$ and the $S3$ models the WER become 22.9 and 22.0 with OOD LM. Finally, it can be seen that at the $S3$ model there is no significant improvement in pseudo-labels. It means no further improvements is expected if more stages are trained afterwards.

5. CONCLUSION

This paper presented multi-stage student training method in combination with ensemble teacher models for improved unsupervised domain adaptation. The student models evaluated on SWBD test sets have shown significant improvement starting from the first stage to the third stage student. Each student was trained using the pseudo-labels generated by the previous teacher models. The progressive improvement at each stage in terms of WER has shown the effectiveness of employing multi-stage training method. Furthermore, it was evaluated that the pseudo-labels with higher WER do not show any advantage of using wav2vec-large model compared to the wav2vec-base. However, as the labels get better in subsequent stages the model based on wav2vec-large gives better performance.

6. ACKNOWLEDGMENT

This work was conducted at the LivePerson centre of Speech and Language Technology at the university of Sheffield which is supported by LivePerson, Inc.

7. REFERENCES

[1] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking evaluation in asr: Are our models robust enough?," *arXiv preprint arXiv:2010.11745*, 2020.

[3] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: An overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33–66, 2020.

[4] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. N. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition.," in *Interspeech*, pp. 892–896, 2018.

[5] S. Khurana, N. Moritz, T. Hori, and J. Le Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6553–6557, IEEE, 2021.

[6] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki, "Exploiting adapters for cross-lingual low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2021.

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

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[9] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition.," in *Interspeech*, pp. 3439–3443, 2016.

[10] H.-G. Kim, H. Na, H. Lee, J. Lee, T. G. Kang, M.-J. Lee, and Y. S. Choi, "Knowledge distillation using output errors for self-attention end-to-end models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6181–6185, IEEE, 2019.

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

[12] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5185–5189, IEEE, 2017.

[13] Z. Meng, J. Li, Y. Gaur, and Y. Gong, "Domain adaptation via teacher-student learning for end-to-end speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 268–275, IEEE, 2019.

[14] H. Zhu, J. Zhao, Y. Ren, L. Wang, and P. Zhang, "Domain adaptation using class similarity for robust speech recognition," *arXiv preprint arXiv:2011.02782*, 2020.

[15] Z.-q. Zhang, Y. Song, J.-s. Zhang, I. McLoughlin, and L.-r. Dai, "Semi-supervised end-to-end asr via teacher-student learning with conditional posterior distribution.," in *INTERSPEECH*, pp. 3580–3584, 2020.

[16] B. Kim, S. Yang, J. Kim, and S. Chang, "Domain generalization on efficient acoustic scene classification using residual normalization," *arXiv preprint arXiv:2111.06531*, 2021.

[17] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers.," in *Interspeech*, pp. 3697–3701, 2017.

[18] Y. Gao, T. Parcollet, and N. D. Lane, "Distilling knowledge from ensembles of acoustic models for joint ctc-attention end-to-end speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 138–145, IEEE, 2021.

[19] S. Li, X. Lu, S. Sakai, M. Mimura, and T. Kawahara, "Semi-supervised ensemble dnn acoustic model training," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5270–5274, IEEE, 2017.

[20] R. Ahmad, M. A. Jalal, M. U. Farooq, A. Ollerenshaw, and T. Hain, "Towards domain generalisation in asr with elitist sampling and ensemble knowledge distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[21] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *arXiv preprint arXiv:2005.09267*, 2020.

[22] J. Rathod, N. Dawalatabad, S. Singh, and D. Gowda, "Multi-stage progressive compression of conformer transducer for on-device speech recognition," *arXiv preprint arXiv:2210.00169*, 2022.

[23] S. Abdou and M. S. Scordilis, "Beam search pruning in speech recognition using a posterior probability-based confidence measure," *Speech Communication*, vol. 42, no. 3, pp. 409–428, 2004.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[25] A. Graves, "Connectionist temporal classification," in *Supervised sequence labelling with recurrent neural networks*, pp. 61–93, Springer, 2012.

[26] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, pp. 28–39, Springer Berlin Heidelberg, 2006.

[27] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE ICASSP*, pp. 5206–5210, 2015.

[29] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *IEEE ICASSP*, vol. 1, pp. 517–520 vol.1, 1992.