ENHANCING AND ADVERSARIAL: IMPROVE ASR WITH SPEAKER LABELS

Wei Zhou[1,2], Haotian Wu[1], Jingjing Xu[1], Mohammad Zeineldeen[1,2], Christoph Lüscher[1,2], Ralf Schlüter[1,2], Hermann Ney[1,2]

[1]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52074 Aachen, Germany
[2]AppTek GmbH, 52062 Aachen, Germany

ABSTRACT

ASR can be improved by multi-task learning (MTL) with domain enhancing or domain adversarial training, which are two opposite objectives with the aim to increase/decrease domain variance towards domain-aware/agnostic ASR, respectively. In this work, we study how to best apply these two opposite objectives with speaker labels to improve conformer-based ASR. We also propose a novel adaptive gradient reversal layer for stable and effective adversarial training without tuning effort. Detailed analysis and experimental verification are conducted to show the optimal positions in the ASR neural network (NN) to apply speaker enhancing and adversarial training. We also explore their combination for further improvement, achieving the same performance as i-vectors plus adversarial training. Our best speaker-based MTL achieves 7% relative improvement on the Switchboard Hub5'00 set. We also investigate the effect of such speaker-based MTL w.r.t. cleaner dataset and weaker ASR NN.

Index Terms— ASR, speaker, multi-task, adversarial

1. INTRODUCTION & RELATED WORK

Multi-task learning (MTL) with additional domain information such as accents [1–6] and noisy conditions [7–9] can improve automatic speech recognition (ASR). The secondary domain-related objectives mainly fall into two categories:

Domain Enhancing Training (DET) aims to enhance domain variance in the ASR neural network (NN). Although this appears contradictory to the ASR objective, it may help the model to learn better transformation towards a domain-aware speech content classification [1–3]. This is usually done via an additional domain discriminator. It shares certain similarity as integrating domain embeddings into the ASR NN, with a major difference of requiring the ASR NN to learn domain discriminative representations itself. Note that this category was often referred to as domain MTL in some literature, but we use the term enhancing to explicitly differentiate it from the next category, which is also one type of MTL.

Domain Adversarial Training (DAT) aims to reduce domain variance in the ASR NN towards a domain-agnostic speech content classification for better generalization [5,7–9], which is consistent with the ASR objective. It also allows domain adaptation with un-transcribed audio from the target domain, where the knowledge transfer from the source domain

is enabled by learning domain-invariant representations [4,6]. The most common way to apply DAT is the gradient reversal layer (GRL) [10], which reverses the gradient from a subsequent domain discriminator, while other approaches such as [11, 12] can also be applied.

DET and DAT are opposite objectives, yet each of them is successfully applied to improve ASR. Thorough studies of their best utilization and possible combination are somewhat missing. Some brief attempts to compare them [4, 13] show that DAT is better than DET. For cross-domain sentiment classification, [14] proposed to apply DET to a self-supervised pretrained language model (LM) so that domain specific features can be more easily removed by subsequent DAT. But the effect and feasibility for ASR are not clear.

In this work, we study how to best apply DET and DAT with speaker labels for the speaker variability problem in ASR. This yields the **speaker enhancing training** and the **speaker adversarial training**, respectively, where the latter is shown to improve ASR [15, 16]. We also propose a novel adaptive GRL for stable and effective adversarial training with negligible tuning effort. To join the benefit of both speaker enhancing and adversarial training, we further investigate their combinations w.r.t. both training pipeline and positions in the ASR NN. We use connectionist temporal classification (CTC) [17] for the primary ASR task, and conduct detailed analysis and experiments on the Switchboard (SWB) [18] and Librispeech (LBS) [19] corpora. Our best speaker-based MTL achieves the same improvement as i-vectors [20] plus adversarial training. We also show some generalization and limitation of such speaker-based MTL.

2. MULTI-TASK LEARNING

With ASR being the primary task, our speaker-based MTL framework can be visualized as Fig. 1. We denote θ_{ASR} as the parameter set of the full ASR NN, which is primarily trained with an ASR objective $\mathcal{L}_{\text{ASR}}$ (in this work $\mathcal{L}_{\text{ASR}} = \mathcal{L}_{\text{CTC}}$). In the following, we define the additional speaker-based training branches in more details.

2.1. Speaker Enhancing Training

Let X denote the acoustic feature sequence of a speech utterance, and D_{spk} denote the corresponding speaker label. The speaker enhancing training aims to boost speaker discrimination from the output of some layer l_1 of the ASR NN (denoted
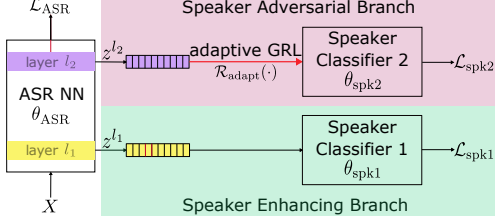
Fig. 1: *ASR + speaker-based multi-task learning framework*

as z^{l_1}). This is done via an additional speaker classifier with parameter set θ_{spk1} to maximize the target speaker posterior based on z^{l_1}, which yields the cross-entropy (CE) loss:

$$\mathcal{L}_{spk1}(\theta_{spk1}, \theta_{ASR}^{l_1}) = -\log P_1(D_{spk}|X; \theta_{spk1}, \theta_{ASR}^{l_1})$$

Here we denote $\theta_{ASR}^{l_1}$ as the relevant part of θ_{ASR} to compute z^{l_1}. The MTL objective can then be obtained as:

$$\mathcal{L}_{spk\text{-enh}}(\theta_{ASR}, \theta_{spk1}) = \mathcal{L}_{ASR}(\theta_{ASR}) + \lambda_1 \cdot \mathcal{L}_{spk1}(\theta_{spk1}, \theta_{ASR}^{l_1})$$

To avoid scale tuning for λ_1, we further apply focal loss [21]:

$$\lambda_1 = [1 - P_1(D_{spk}|X)]^{\beta_{focal}}$$

where β_{focal} is usually omitted (set to 1).

2.2. Speaker Adversarial Training

The speaker adversarial training aims to force speaker-invariant representations from the output of some layer l_2 of the ASR NN (denoted as z^{l_2}). Similarly, another speaker classifier with parameter set θ_{spk2} is introduced to maximize the target speaker posterior based on z^{l_2}. The key difference here is the usage of the GRL $\mathcal{R}(\cdot)$ [10]:

$$\mathcal{R}(z^{l_2}) = z^{l_2}, \quad \frac{d\mathcal{R}}{dz^{l_2}} = -\mathbf{I}$$

where $\mathbf{I}$ is an identity matrix. The GRL simply passes the input in forwarding and reverses the gradient with a negative sign in backpropagation. The speaker adversarial training is then achieved by applying $\mathcal{R}(\cdot)$ with the following CE loss:

$$\mathcal{L}_{spk2}(\theta_{spk2}, \mathcal{R}(\theta_{ASR}^{l_2})) = -\log P_2(D_{spk}|X; \theta_{spk2}, \theta_{ASR}^{l_2})$$

where $\theta_{ASR}^{l_2}$ denotes the relevant part of θ_{ASR} to compute z^{l_2}. In this case, θ_{spk2} and $\theta_{ASR}^{l_2}$ lead to competing update, where the latter is reversely trained to produce confusing z^{l_2} that reduces speaker discrimination.

2.2.1. Standard Approach

The speaker adversarial MTL loss is commonly applied as:

$$\mathcal{L}_{spk\text{-adv}}(\theta_{ASR}, \theta_{spk2}) = \mathcal{L}_{ASR}(\theta_{ASR}) + \lambda_2 \cdot \mathcal{L}_{spk2}(\theta_{spk2}, \mathcal{R}(\theta_{ASR}^{l_2}))$$

which allows to compute all gradient updates including the competing ones in a single backpropagation. Note that we avoid the usual min-max optimization expression here to pose a more straightforward interpretation of the GRL operation.

In this case, the same $\lambda_2 > 0$ scales the opposite update of θ_{spk2} and $\theta_{ASR}^{l_2}$ simultaneously, which can be difficult to tune in general. A large λ_2 may cause unstable training (at the beginning), especially when θ_{spk2} is from-scratch ini-

tialized and produces noisy gradient. Yet a small λ_2 may lead to insufficient adversarial training effect eventually, especially when θ_{spk2} needs to quickly catch up with (pretrained) θ_{ASR}. [10] proposed to schedule an increasing λ_2 only to the GRL, and [6] proposed to pretrain the domain classifier with the frozen primary NN before DAT. Both ease the problem with some additional complexity. But the optimal scaling still requires careful tuning and evaluation on the validation set, which can be time and resource consuming.

2.2.2. Proposed Adaptive GRL

We propose a novel adaptive GRL $\mathcal{R}_{adapt}(\cdot)$ for stable and effective adversarial training with negligible tuning effort:

$$\mathcal{L}_{spk\text{-adv}}^{adapt}(\theta_{ASR}, \theta_{spk2}) = \mathcal{L}_{ASR}(\theta_{ASR}) + \mathcal{L}_{spk2}(\theta_{spk2}, \mathcal{R}_{adapt}(\theta_{ASR}^{l_2}))$$

$$\mathcal{R}_{adapt}(z^{l_2}) = z^{l_2}, \quad \frac{d\mathcal{R}_{adapt}}{dz^{l_2}} = -\lambda_{adapt} \cdot \mathbf{I}$$

$$\lambda_{adapt} = P_2(D_{spk}|X)^{\beta_{adapt}}$$

The scaling on $\mathcal{L}_{spk2}$ is removed to allow fast training of the discriminator θ_{spk2} all the time. More importantly, we scale the reverse update of $\theta_{ASR}^{l_2}$ proportionally to the target speaker posterior from the discriminator. Effectively, this allows more reverse update for higher speaker variance in z^{l_2}, and less update for lower variance, which is a desired behavior towards speaker-invariant representations. This is the key difference to the standard approach, which does exactly the opposite. The strength of such adaptive adversarial training can be adjusted by β_{adapt}. We empirically find that $\beta_{adapt} = 1$ works generally well, although it can be further adjusted w.r.t. the cardinality of D_{spk} and the height of l_2, which might affect the general magnitude of P_2. Note that for a mini-batch of multiple utterances, we simply take the average P_2 for λ_{adapt}.

The proposed $\mathcal{R}_{adapt}$ also ensures a stable training. At the beginning of training, a rather small $P_2(D_{spk}|X; \theta_{spk2}, \theta_{ASR}^{l_2})$ can be expected, which yields a rather large $\mathcal{L}_{spk2}$. This leads to an aggressive training of θ_{spk2}, while a very small λ_{adapt} forbids much reverse gradient to update $\theta_{ASR}^{l_2}$. After θ_{spk2} quickly catches up, an increasing λ_{adapt} allows more reverse gradient to update $\theta_{ASR}^{l_2}$, which gradually reduces the speaker variance in z^{l_2}. This procedure will eventually saturate with little speaker variance left and a fluctuating $\mathcal{L}_{spk2}$ due to the competing update of θ_{spk2} and $\theta_{ASR}^{l_2}$.

2.3. Enhancing + Adversarial

Recent studies [22, 23] on self-supervised speech models show that those representations from lower-middle layers are more suitable for speaker-related tasks such as speaker identification (SID), while those from higher layers are more suitable for content-related tasks such as ASR. Inspired by these findings, we hypothesize that this might be a natural preference of DNN-based speech models, which reveals a possible combination of the two opposite objectives. Specifically, we propose to apply speaker enhancing training to lower layers (l_1) and speaker adversarial training to higher layers (l_2), which yields the following MTL objective:
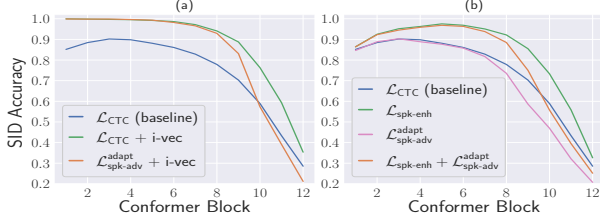
Fig. 2: *Speaker identification (SID) accuracy analysis of each conformer block of various SWB CTC models*

$$\mathcal{L}_{\text{spk-enh-adv}}(\theta_{\text{ASR}}, \theta_{\text{spk1}}, \theta_{\text{spk2}}) = \mathcal{L}_{\text{ASR}}(\theta_{\text{ASR}}) +$$

$$\lambda_1 \cdot \mathcal{L}_{\text{spk1}}(\theta_{\text{spk1}}, \theta_{\text{ASR}}^{l_1}) + \mathcal{L}_{\text{spk2}}(\theta_{\text{spk2}}, \mathcal{R}_{\text{adapt}}(\theta_{\text{ASR}}^{l_2}))$$

The goal is to enhance speaker awareness at lower layers and remove residual speaker variance at higher layers, both of which may jointly benefit the primary ASR task.

3. EXPERIMENTS

3.1. Setups

We conduct most experiments on the 300h SWB corpus [18] containing 520 speakers, and perform some generalization investigation on the 960h LBS corpus [19] containing 2338 speakers. For ASR, we use a phoneme-based CTC model, which contains a VGG network followed by 12 conformer [24] blocks and a softmax layer (details see [25]). All speaker classifiers adopt a simple MLP attention-based pooling followed by a softmax layer for the speaker posterior. With only 1 hidden layer of 512 units for the MLP attention, each speaker classifier contains only 0.5M parameters. For z^{l_2}, we always use the conformer block output, while for z^{l_1}, we empirically find that using the conformer block output before layer normalization gives better results. Both β_{focal} and β_{adapt} are set to 1. We use gammatone features [26] and specaugment [27] in all training. For efficiency, we always start from a seed CTC model and reset learning rate (LR) with a constant + linear decay policy to continue training with various objectives (SWB: 25-30 epochs; LBS: 10-12 epochs). In this case, the speaker classifiers are always from-scratch initialized. For fair comparison, the baseline model is also continue-trained with $\mathcal{L}_{\text{CTC}}$ only. By default, we apply Viterbi decoding with a 4gram word-level LM. For SWB, we use Hub5'00 and Hub5'01 as dev and test sets, respectively.

3.2. Pre-analysis with I-vectors

Speaker embeddings such as i-vectors [20] are known to contain rich speaker information and improve ASR performance [28,29]. To have a better idea of where to apply which speaker training, we firstly try to analyze how much speaker variance is contained at each conformer block of the ASR NN, before and after adding such speaker discriminative representations. We analyze this by performing SID task, where we train a speaker classifier for each conformer block output of a frozen ASR model. This is done on the SWB corpus with a random 5% split for evaluation. We apply the Weighted-Simple-Add method as [29] to integrate i-vectors in the 1st conformer block during the continue-training phase.

The word error rate (WER) results of the baseline and

Table 1: *WER [%] results of various CTC models on Hub5'00 including the Switchboard (SWB) and CallHome (CH) subsets, decoded with a 4gram LM.*

Model Train	Conformer Block		Hub5'00		
	l_1	l_2	SWB	CH	$\sum$
$\mathcal{L}_{\text{CTC}}$ (baseline)			7.6	17.3	12.4
$\mathcal{L}_{\text{CTC}}$ + i-vec	-		7.7	16.1	11.9
$\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$ + i-vec	-	9	7.5	16.0	11.8
		10	7.4	15.8	**11.6**
		11	7.6	16.4	12.0
		12	7.6	16.1	11.9
$\mathcal{L}_{\text{spk-enh}}$	1	-	7.8	16.7	12.3
	2		7.6	16.6	12.1
	3		7.6	16.8	12.2
	4		7.7	16.5	12.1
	5		7.7	16.2	**12.0**
	6		7.7	16.6	12.2
	7		7.8	17.1	12.4
$\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$	-	7	7.9	17.1	12.5
		8	7.7	16.8	12.3
		9	7.7	16.4	**12.1**
		10	7.7	16.9	12.3
		11	7.7	16.7	12.2
		12	7.6	16.8	12.2
$\mathcal{L}_{\text{spk-adv}}$, $\lambda_2 = 0.1$		9	7.8	16.9	12.4
$\lambda_2 = 0.5$			7.7	16.7	12.2
$\lambda_2 = 1.0$			7.6	16.9	12.3
$\mathcal{L}_{\text{spk-enh-adv}}$	**5**	9	7.7	16.8	12.3
		10	7.6	16.6	12.1
$\mathcal{L}_{\text{spk-enh}} + \mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$		**9**	7.5	15.7	**11.6**
		10	7.5	16.0	11.8

i-vectors based CTC models are shown in Tab. 1, where i-vectors give 4% relative improvement. The corresponding SID analysis of these two models are shown in Fig. 2(a), where i-vectors increase the SID accuracy of all conformer blocks. The most accuracy boost occurs at blocks 1-6 (almost 100%), which gives a clear indication for speaker enhancing training. The increment becomes less for higher blocks with lower accuracy, which matches the ASR objective. However, the ASR objective does not explicitly force a low residual speaker variance, which can also be seen from the final block's SID accuracy. This seems to indicate the necessity of speaker adversarial training. As a result, we apply i-vectors together with our $\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$ for higher conformer blocks. The results are also shown in Tab. 1, where the best $l_2 = 10$ further improves the performance. We also show the SID analysis of this model in Fig. 2(a), which gives a clear proof of concept for our proposal in Sec. 2.3.

3.3. Speaker Labels Only

We then verify those findings from the above analysis on the speaker-based MTL. Besides the study of best utilization, the goal is also to achieve similar improvement without the use of any speaker embeddings, which usually introduce extra complexity in run time.

3.3.1. Speaker Enhancing Training

We apply $\mathcal{L}_{\text{spk-enh}}$ for conformer blocks 1-7, whose results are shown in Tab. 1. All $l_1 < 7$ improve the baseline. The best
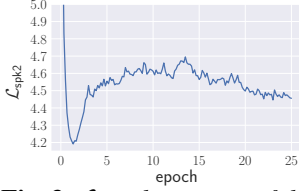
Fig. 3: $\mathcal{L}_{spk2}$ loss curve of the SWB CTC model trained with $\mathcal{L}_{spk\text{-}adv}^{adapt}$ ($l_2 = 9$)
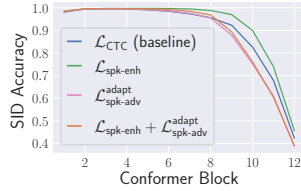

Fig. 4: *SID accuracy ayalysis of each conformer block of LBS CTC models*

Table 2: *WER [%] results of best SWB CTC models under various objectives and corresponding LBS CTC models under the same settings, all decoded with transformer LM.*

Model Train	Hub5'00			Hub5'01	LBS dev		LBS test	
	SWB	CH	$\sum$		clean	other	clean	other
$\mathcal{L}_{\text{CTC}}$ (baseline)	6.8	16.1	11.5	10.5	1.9	4.4	2.3	4.9
$\mathcal{L}_{\text{spk-enh}}$	6.6	15.2	10.9	10.3	1.9	4.4	2.3	4.8
$\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$	6.9	15.1	11.0	10.4	1.9	4.3	2.3	4.8
$\mathcal{L}_{\text{spk-enh}} + \mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$	**6.6**	**14.8**	**10.7**	**10.2**	**1.9**	**4.3**	**2.2**	**4.7**

$l_1 = 5$ achieves almost the same improvement as i-vectors. We also perform the same SID analysis on this model, which is shown in Fig. 2(b). A very similar accuracy boost pattern as i-vectors is observed for blocks ≥ 5, although the effect is still weaker than i-vectors. This indicates that the NN-learned speaker discriminative representations is not as powerful as those explicitly extracted speaker embeddings, which may contain additional channel information.

3.3.2. Speaker Adversarial Training

We then apply $\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$ for conformer blocks 7-12, whose results are also shown in Tab. 1. All $l_2 > 7$ improve the baseline, where $l_2 = 9$ gives the best improvement. We did not spend much tuning effort here, but we observe that ($l_2 = 12$, $\beta_{\text{adapt}} = 0.5$) can also achieve this result. For comparison, we also evaluate the standard $\mathcal{L}_{\text{spk-adv}}$ for $l_2 = 9$ and $\lambda_2 \in \{0.1, 0.5, 1.0\}$. As shown in Tab. 1, they are all worse than our proposed adaptive adversarial training.

Again, we show the best model's SID analysis in Fig. 2(b), where a clear accuracy drop is observed for blocks ≥ 8. Together with the SID analysis for $\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$ + i-vectors, it seems that the adversarial training only has a local effect to reduce speaker variance of conformer blocks $\geq l_2 - 1$, where lower blocks are rarely affected. Additionally, we also show the training $\mathcal{L}_{spk2}$ loss curve of this model in Fig. 3, which verifies the stable training behavior described in Sec. 2.2.2.

3.3.3. Enhancing + Adversarial

To combine the benefit of both speaker enhancing and adversarial training, we directly apply $\mathcal{L}_{\text{spk-enh-adv}}$ to jointly train them in one run, where the best $l_1 = 5$ and $l_2 = 9/10$ from previous experiments are used. As shown in Tab. 1, such direct joint training does not lead to further improvement over the individual best models. The results also do not get better after some careful tuning of LR schedule and longer training.

Based on the previous SID analysis, we suspect that the two objectives may have conflicting effect around l_2, as they try to increase and decrease speaker discrimination simultaneously. As a result, we propose to sequentially apply speaker enhancing and adversarial training at lower and higher layers, respectively. More precisely, we firstly train a speaker-aware ASR model using $\mathcal{L}_{\text{spk-enh}}$, and then fine-tune it with $\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$, which utilizes its local effect to gradually remove the residual speaker variance at higher layers. We denote this as $\mathcal{L}_{\text{spk-enh}} + \mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$. As shown in Tab. 1, this approach achieves further improvement, where the best ($l_1 = 5, l_2 = 9$) gives the

same performance as the best model of $\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$ + i-vectors. Their SID analysis in Fig. 2 also show very similar pattern.

3.4. Generalization and Limitation

We evaluate the best model of each speaker-based MTL with transformer LM 1-pass decoding. As shown in Tab. 2, the improvements carry over to both Hub5'00 and Hub5'01 sets, where the nosier CallHome (CH) subset benefits the most.

We then apply the best SWB settings to LBS, whose WER results and corresponding SID analysis are shown in Tab. 2 and Fig. 4, respectively. Surprisingly, the baseline CTC model already achieves almost 100% SID accuracy at lower conformer blocks. Consistent improvements are still obtained for all speaker-based MTL, but the overall effect is much smaller for such clean data with matched conditions.

To investigate the effect of weaker NN on speaker-based MTL, we replace conformer with a 6×512 bidirectional long short-term memory [30] (BLSTM) for the SWB experiments. For $\mathcal{L}_{\text{spk-enh}}$, we get no improvement over the baseline, which indicates its strong dependency on the NN's capability to learn speaker awareness from the input X. For $\mathcal{L}_{\text{spk-adv}}^{\text{adapt}}$, we only get minor improvement, similar as reported in [15].

4. CONCLUSION

In this work, we showed how to best apply speaker enhancing training and speaker adversarial training in a multi-task learning (MTL) framework to improve conformer-based ASR. We proposed a novel adaptive GRL for stable and effective adversarial training without tuning effort, which outperforms the standard approach. Our detailed analysis and experimental verification suggest to apply the enhancing training at lower layers of the ASR NN to enhance speaker awareness, and the adversarial training at higher layers to remove speaker variance matching the ASR objective. The two opposite objectives at different layers can also be sequentially combined for further improvement, achieving the same performance as i-vectors plus adversarial training. This improves the baseline CTC model by 7% relative on the Switchboard Hub5'00 set. The effect of such speaker-based MTL is smaller on the cleaner Librispeech dataset, and is limited by weaker NN such as small BLSTM.

5. REFERENCES

[1] Abhinav Jain, Minali Upreti, and Preethi Jyothi, "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning," in *Proc. Interspeech*, 2018, pp. 2454–2458.

[2] Thibault Viglino, Petr Motlícek, and Milos Cernak, "End-to-End Accented Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2140–2144.

[3] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, "Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5989–5993.

[4] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain Adversarial Training for Accented Speech Recognition," in *Proc. ICASSP*, 2018, pp. 4854–4858.

[5] Aditay Tripathi, Aanchan Mohan, Saket Anand, and Maneesh Singh, "Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5959–5963.

[6] Nilaksh Das, Sravan Bodapati, Monica Sunkara, Sundararajan Srinivasan, and Duen Horng Chau, "Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning," in *Proc. Interspeech*, 2021, pp. 1314–1318.

[7] Yusuke Shinohara, "Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition," in *Proc. Interspeech*, 2016, pp. 2369–2372.

[8] Dmitriy Serdyuk, Kartik Audhkhasi, Philemon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, "Invariant Representations for Noisy Speech Recognition," 2016, http://arxiv.org/abs/1612.01928.

[9] Zhong Meng, Zhuo Chen, Vadim Mazalov, Jinyu Li, and Yifan Gong, "Unsupervised Adaptation with Domain Separation Networks for Robust Speech Recognition," in *IEEE ASRU*, 2017, pp. 214–221.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, pp. 59:1–59:35, 2016.

[11] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain Separation Networks," in *Proc. NeurIPS*, 2016, pp. 343–351.

[12] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko, "Adversarial Dropout Regularization," in *Int. Conf. on Learning Representations (ICLR)*, 2018.

[13] Tomohiro Tanaka, Ryo Masumura, Hiroshi Sato, Mana Ihori, Kohei Matsuura, Takanori Ashihara, and Takafumi Moriya, "Domain Adversarial Self-Supervised Speech Representation Learning for Improving Unknown Domain Downstream Tasks," in *Proc. Interspeech*, 2022, pp. 1066–1070.

[14] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao, "Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis," in *Proc. ACL*, 2020, pp. 4019–4028.

[15] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," in *Proc. Interspeech*, 2017, pp. 132–136.

[16] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gong, and Biing-Hwang Juang, "Speaker-Invariant Training Via Adversarial Learning," in *Proc. ICASSP*, 2018, pp. 5969–5973.

[17] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, pp. 369–376.

[18] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCH-BOARD: Telephone Speech Corpus for Research and Development," in *Proc. ICASSP*, 1992, vol. 1, pp. 517–520.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[20] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[21] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, "Focal Loss for Dense Object Detection," in *IEEE International Conference on Computer Vision ICCV*, 2017, pp. 2999–3007.

[22] Heng-Jui Chang, Shu-Wen Yang, and Hung-yi Lee, "Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert," in *Proc. ICASSP*, 2022, pp. 7087–7091.

[23] Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, and Tomohiro Tanaka, "Deep versus Wide: An Analysis of Student Architectures for Task-Agnostic Knowledge Distillation of Self-Supervised Speech Models," in *Proc. Interspeech*, 2022, pp. 411–415.

[24] Anmol Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[25] Wei Zhou, Wilfried Michel, Ralf Schlüter, and Hermann Ney, "Efficient Training of Neural Transducer for Speech Recognition," in *Proc. Interspeech*, 2022, pp. 2058–2062.

[26] Ralf Schlüter, Ilja Bezrukov, Hermann Wagner, and Hermann Ney, "Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition," in *Proc. ICASSP*, 2007, pp. 649–652.

[27] Barret Zoph, Chung-Cheng Chiu, Daniel S. Park, Ekin Dogus Cubuk, Quoc V. Le, William Chan, and Yu Zhang, "SpecAugment: A Simple Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[28] Markus Kitza, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Cumulative Adaptation for BLSTM Acoustic Models," in *Proc. Interspeech*, 2019, pp. 754–758.

[29] Mohammad Zeineldeen, Jingjing Xu, Christoph Lüscher, Ralf Schlüter, and Hermann Ney, "Improving the Training Recipe for a Robust Conformer-based Hybrid Model," in *Proc. Interspeech*, 2022, pp. 1036–1040.

[30] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.