MADI: INTER-DOMAIN MATCHING AND INTRA-DOMAIN DISCRIMINATION FOR CROSS-DOMAIN SPEECH RECOGNITION

Jiaming Zhou¹

Shiwan Zhao*

Ning Jiang²

Yong Oin^{1†}

¹ Nankai University, Tianjin, China ² Mashang Consumer Finance Co., Ltd.

ABSTRACT

End-to-end automatic speech recognition (ASR) usually suffers from performance degradation when applied to a new domain due to domain shift. Unsupervised domain adaptation (UDA) aims to improve the performance on the unlabeled target domain by transferring knowledge from the source to the target domain. To improve transferability, existing UDA approaches mainly focus on matching the distributions of the source and target domains globally and/or locally, while ignoring the model discriminability. In this paper, we propose a novel UDA approach for ASR via inter-domain MAtching and intra-domain DIscrimination (MADI), which improves the model transferability by fine-grained inter-domain matching and discriminability by intra-domain contrastive discrimination simultaneously. Evaluations on the Libri-Adapt dataset demonstrate the effectiveness of our approach. MADI reduces the relative word error rate (WER) on cross-device and cross-environment ASR by 17.7% and 22.8%, respectively.

Index Terms— speech recognition, domain adaptation, transferability, discriminability

1. INTRODUCTION

Recent years have witnessed great progress in end-to-end automatic speech recognition (ASR) based on deep learning methods [1], which rely on large-scale labeled datasets and assume that training and testing data come from the same distribution. Nevertheless, when the models are trained on one domain (source domain) and tested on another domain (target domain), the performance degrades severely due to cross-domain distribution shift (domain shift). The causes of domain shift include variabilities of the acoustic environment, device, accent, and so on. Collecting sufficient labeled data for each target domain to train a good ASR model is expensive and time-consuming.

Unsupervised domain adaptation (UDA) has been proposed to improve the ASR performance on the unlabeled target domain by leveraging the label-rich source domain. To transfer knowledge from the source to the target domain, previous work mainly focuses on matching the distributions of the source and target domains by learning domain-invariant representations. Generative adversarial nets (GAN) [2, 3] and domain adversarial learning technique [4, 5, 6, 7, 8] have shown to be effective for global domain matching. To name a few, Chen et al. [2] attempt to disentangle accent-specific and accent-invariant characteristics to build a unified end-to-end ASR system based on GAN. Sun et al. [4] propose domain adversarial training (DAT) to encourage the model to learn domain-invariant representations. Discrepancy-based methods, such as maximum mean discrepancy (MMD) [9] and correlation (CORAL) [10], have recently been used to minimize feature distribution discrepancy between domains.

Guoging Zhao²

More recently, local domain matching approaches have become popular for fine-grained distribution matching, which aligns the distributions of the relevant subdomains across different domains. Hu et al. [11] propose subdomain distribution matching to extract domain-invariant embeddings for speaker verification. In CMatch [12], a character-level distribution matching method is adopted to address domain shift. The inter-domain matching, either globally or locally, improves the model transferability. However, simply pushing the source and target domains together may compromise the discriminability of the model in the target domain [13, 14].

To address the aforementioned issue, in this paper, we propose MADI, a novel UDA approach for ASR via interdomain matching and intra-domain discrimination. With fine-grained inter-domain matching, the proposed method improves the model transferability, while with intra-domain contrastive discrimination, we enhance the model discriminability in the target domain. Specifically, our framework contains two main components for domain adaptation (see figure 1). Firstly, inspired by CMatch [12], we employ an inter-domain matching component that matches the characterlevel distributions between the labeled source domain and unlabeled target domain. Secondly, motivated by the success of contrastive learning [15, 16], we generate augmented unlabeled target data and then propose an intra-domain discrimination component to ensure that the centroids of the same characters are pulled closer, while the centroids of dif-

^{*} Independent researcher.

[†]Corresponding author. This work was supported in part by NSF China (Grant No. 62271270), Tianjin Media Computing Center, Big Data Institute of Nankai University, and Mashang Consumer Finance.



Fig. 1. Overview of our MADI framework. For source samples with ground-truth labels and (augmented) target samples with pseudo labels, we use encoders to extract features, and then employ the CTC decoder to assign labels to the encoded frames. Then we compute the two adaptation losses: \mathcal{L}_{MA} for inter-domain matching and \mathcal{L}_{DI} for intra-domain discrimination. The former matches the character-level distributions between the source and target domains. The latter employs the contrastive learning method to push the centroids of different characters away from each other in the target domain. For concise, the joint CTC-Attention loss \mathcal{L}_{ASR} is omitted in the figure.

ferent characters are pushed apart in the target domain. This facilitates learning discriminative representations for the unlabeled target domain. The above two components are jointly optimized using the well-defined MMD [9] and normalized temperature-scaled cross-entropy loss (NT-Xent) [15].

We note that cross-domain contrastive learning approaches [17, 18] from the computer vision domain are similar to our work, which performs inter-domain alignment and intradomain discrimination by deliberately constructing positive and negative pairs. Directly adopting such contrastive learning approaches at the frame level results in an explosion of computational complexity, while at the character-prototype level it leads to sub-optimal performance (see Section 4).

We conduct extensive experiments on the Libri-Adapt dataset [19]. The results demonstrate the effectiveness of our approach. MADI outperforms the state-of-the-art UDA methods and achieves a relative performance improvement of 17.7% and 22.8% word error rate (WER) on cross-device and cross-environment ASR, respectively.

2. OUR METHOD

UDA for ASR aims to exploit labeled source data (X_S, Y_S) to improve the ASR performance on unlabeled target data X_T . Our key idea is to match the character-level distributions between the source and target domains to enhance the model transferability and push the features of different characters apart to improve the model discriminability in the target domain. The overall structure of the proposed adaptation method, inter-domain matching and intra-domain discrimination (MADI), is shown in figure 1. Before describing the two adaptation components in detail, we briefly introduce the basic ASR model used in our framework.

2.1. Basic ASR Model

We build a joint CTC-Attention model following open-source Wenet [20], which consists of three parts: shared encoder, CTC decoder, and attention decoder. The loss of ASR is as follows:

$$\mathcal{L}_{ASR}(X,Y) = \lambda \mathcal{L}_{CTC}(X,Y) + (1-\lambda)\mathcal{L}_{ATT}(X,Y),$$
(1)

where X and Y are the acoustic input and corresponding labels, respectively. \mathcal{L}_{CTC} is the CTC loss, and \mathcal{L}_{ATT} is the attention loss. The hyperparameter λ balances the two losses.

2.2. Inter-domain Matching

Since our inter-domain matching is based on character level, we first assign labels to frames, which is time-consuming. We follow CMatch [12] to achieve efficient and accurate frame-level label assignment. With the pre-trained model on (X_S, Y_S) , we obtain the CTC [21] pseudo label of the *n*-th frame by

$$\hat{Y}_n = \arg\max_{Y_n} P_{CTC}(Y_n | X_n).$$
⁽²⁾

After acquiring labels for each encoded frame, the conditional distributions P(Y|X) of each character in both the source and target domains could be obtained. We then adopt the widely used MMD distance [9] to match the conditional distributions between the same characters across domains. The inter-domain matching loss \mathcal{L}_{MA} is as follows:

$$\mathcal{L}_{MA} = \frac{1}{N} \sum_{i}^{N} MMD(\mathcal{H}_k, X_S^{C_i}, X_T^{C_i}), \qquad (3)$$

where N is the total number of characters. C_i means the *i*-th character of the symbol set C. \mathcal{H}_k is the reproducing kernel Hilbert space, and k is the Gaussian kernel function we

adopted. The model is trained to minimize \mathcal{L}_{MA} to match the distributions of the same characters across domains.

2.3. Intra-domain Discrimination

We employ the contrastive learning approach [15] in the target domain to improve the model discriminability. Given target domain data X_T , we first generate an augmented version X_{aug} by pitch randomization, reverberation, and temporal masking. And then, similar to inter-domain matching, we achieve frame-level label assignment using CTC pseudo labels for both X_T and X_{aug} . Adopting contrastive learning at the frame level will result in an explosion of computational complexity, so we apply the contrastive learning on the character prototypes by computing the centroids for 2*N symbols in a batch of X_T and X_{aug} . The character centroids of the same symbol form positive pairs $(\widetilde{X}_T^{C_i}, \widetilde{X}_{aug}^{C_i})$, and the rest ones from X_T and X_{aug} are counted as negative pairs. We attempt to keep positive pairs together and push negative pairs apart in the batch by minimizing the modified NT-Xent loss [15]. The intra-domain discrimination loss \mathcal{L}_{DI} is defined as:

$$\mathcal{L}_{DI}(\widetilde{X}_{T}^{C_{i}}, \widetilde{X}_{aug}^{C_{i}}) = -log \frac{\psi(\widetilde{X}_{T}^{C_{i}}, \widetilde{X}_{aug}^{C_{i}})}{\psi(\widetilde{X}_{T}^{C_{i}}, \widetilde{X}_{aug}^{C_{i}}) + \sum_{d \in \{T, aug\}}^{j \neq i} \psi(\widetilde{X}_{T}^{C_{i}}, \widetilde{X}_{d}^{C_{j}})},$$
(4)

where $1 \leq i, j \leq N$, and $\psi(a, b) = exp(sim(f(a), f(b))/\tau)$. $sim(u, v) = \frac{u^T v}{\|u\|_2 \|v\|_2}$ denotes the cosine similarity of u and v. f() means features extracted by the encoder, and τ is the temperature hyperparameter. Note that \mathcal{L}_{DI} is the average of $\mathcal{L}_{DI}(\widetilde{X}_T^{C_i}, \widetilde{X}_{aug}^{C_i})$ and $\mathcal{L}_{DI}(\widetilde{X}_{aug}^{C_i}, \widetilde{X}_T^{C_i})$ for all positive pairs.

2.4. Overall Loss

The overall loss function includes the joint CTC-Attention loss, the inter-domain matching loss, and the intra-domain discrimination loss, which is defined as follows:

$$\mathcal{L} = \mathcal{L}_{ASR} + \alpha * \mathcal{L}_{MA} + \beta * \mathcal{L}_{DI}, \tag{5}$$

where α and β are hyperparameters to tradeoff the impact of \mathcal{L}_{MA} and \mathcal{L}_{DI} .

3. EXPERIMENTAL SETUP

3.1. Dataset

Our experiments are conducted on the Libri-Adapt dataset [19] which is derived from Librispeech-clean-100. The Libri-Adapt provides 72 different domains for domain adaptation study, which is recorded under 4 background noise conditions from 3 speaker accents on 6 different embedded microphones. In this work, we focus on cross-device and cross-environment adaptation, i.e., source-domain and target-domain data are recorded by different devices or in different environments.

Recent research [22] has shown that the variabilities of microphones across different devices significantly influence their outputs. In this paper, we employ 3 parts of the Libriadapt dataset for cross-device experiments including Matrix Voice (M), Respeaker (R), and PlayStation Eye (P). Matrix Voice and Respeaker are circular 7-channel microphone arrays integrated with acoustic signal processing algorithms while Playstation-Eye is a 4-channel microphone for voice interactive games. For cross-environment ASR adaptation, we select clean Respeaker as the source domain and 3 types of background noise including Rain, Wind, and Laughter as different target domains. In our experiments, we do not use any labels from the target domain during training and randomly split 10% utterances in the source domain as the validation set.

3.2. Baselines

The following methods are considered for comparison:

- **SO**: The source-only (SO) method trains the ASR model on the source domain and directly applies it to the target domain without adaptation.
- **DAT** [4]: Domain adversarial training (DAT) is a popular UDA method which adversarially trains a discriminator and an encoder to encourage the encoder to learn domain-invariant features. We re-implemented DAT with a domain discriminator consisting of fully-connected linear layers.
- **CMatch** [12]: It is a character-level distribution matching method, which employs CTC pseudo labels to achieve frame-level label assignment and then reduces the characterlevel distribution divergence between the source and target domains using MMD. We also re-implement CMatch.
- **CDCL**: Cross-domain contrastive learning (CDCL) approaches are popular in the computer vision domain [17, 18]. We implement the idea at the character prototype level by considering the centroids of the same characters from different domains as positive pairs and the centroids of different characters from both domains as negative pairs.

3.3. Implementation Details

For fair comparison, we implement all baselines and our method based on Wenet [20] codebase. All experiments use 80-dimensional log Mel-filter banks (FBANK) features with a 25ms window and a 10ms shift. The underlying transformer model has 12 encoder layers and 6 decoder layers. Both of them have 4 attention heads and 2048 linear units. The CTC loss weight λ is set to 0.3 following [20]. The hyperparameters α and β are set to 5 in our method. The temperature τ is 0.1 during our experiment. The training data of the target domain is augmented using the open-source tool WavAugment [23] by pitch randomization, reverberation, and temporal

lard ASR
WER
23.74
20.77
22.82
22.44

masking. When training, we filter out utterances over 17.5s. We employ the learning rate from 1×10^{-3} to 8×10^{-3} and adam optimizer with a learning rate schedule including 25,000 warm-up steps. Beam size is 10 for decoding. The batch size is 64 and the epoch is 150/180 for cross-device/environment models. The output dimension is 31 consisting of 26 letters and 5 symbols. The attention-rescoring mode we adopt at the testing time always keeps the best performance among the 4 decoding methods provided by the model.

4. RESULTS

We first report the WER results of the standard ASR in Table 1 for comparison. The standard ASR is an in-domain model with training and testing data from the same domain.

4.1. Cross-device Adaptation

The main results of cross-device ASR are reported in Table 2. The task name indicates the source and target domains, e.g., $M \rightarrow P$ denotes the source domain Matrix Voice (M) and the target domain Playstation Eye (P).

Firstly, we observe that the performance of SO is severely degraded due to domain mismatch. Both DAT and CMatch improve performance on all tasks through inter-domain matching. CMatch outperforms DAT, indicating that fine-grained local domain alignment is superior to global alignment. Secondly, CDCL is inferior to CMatch, although CDCL attempts to align positive pairs across domains and separate negative pairs at the same time. The reason is that the inter-domain matching ability of CDCL at the character prototype level is weaker than that of MMD used in CMatch. Thirdly, MADI achieves the lowest WER on 5 of the 6 tasks and the best average performance. MADI significantly outperforms SO by 17.7% relatively, which demonstrates the effectiveness of our approach by improving the transferability and discriminability simultaneously.

To further demonstrate the ability of MADI to enhance the model discriminability, we visualize the feature distributions of CMath and MADI in figure 2. We observe that the centroids of the same characters from different domains are well aligned in CMatch while the distances between different characters are somewhat close. By applying contrastive discrimination in the target domain, characters in MADI are pushed away from each other, indicating the improvement of the model discriminability.



Fig. 2. Compared to CMatch, feature centers of characters are more spread out in MADI. Dots in red and blue indicate the source and target domains, respectively.

Table 2. WER on cross-device ASR						
Task	SO	DAT	CMatch	CDCL	MADI	
$M \rightarrow P$	23.94	21.92	20.28	21.53	20.25	
$M {\rightarrow} R$	26.43	23.58	22.79	24.49	22.61	
$P \rightarrow M$	28.97	25.02	23.91	24.99	23.41	
$P \rightarrow R$	23.54	22.64	20.25	22.06	19.7	
$R{\rightarrow}M$	34.95	28.34	28.68	29.45	27.27	
$R{\rightarrow}P$	22.7	21.84	18.82	20.53	18.89	
Average	26.76	23.89	22.46	23.84	22.02	

Table 3. WER on cross-environment ASR							
Task	SO	DAT	CMatch	MADI			
Rain	33.06	33.64	26.24	25.82			
Wind	26.19	27.17	21.40	21.06			
Laughter	31.12	28.52	23.48	22.91			
Average	30.12	29.78	23.71	23.26			

4.2. Cross-environment Adaptation

The cross-environment ASR results are shown in Tabel 3. We also observe that MADI outperforms DAT and CMatch. Moreover, compared to SO trained with Respeaker in the clean environment, MADI reduces relative WER by 22.8%, indicating its effectiveness for cross-environment adaptation. Note that our re-implementation of CMatch performs better than what the paper [12] reports.

5. CONCLUSION

In this paper, we propose an unsupervised cross-domain ASR adaptation method via inter-domain matching and intradomain discrimination. Our approach improves the model transferability and discriminability simultaneously. Experimental results on the Libri-Adapt dataset demonstrate the effectiveness of our approach.

6. REFERENCES

- [1] Jinyu Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, April 2022.
- [2] Yi-Chen Chen, Zhaojun Yang, Ching-Feng Yeh, Mahaveer Jain, and Michael L. Seltzer, "Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition," *ICASSP*, 2019.
- [3] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *CVPR*, 2018, pp. 1498– 1507.
- [4] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain adversarial training for accented speech recognition," *ICASSP*, 2018.
- [5] Nilaksh Das, Sravan Bodapati, Monica Sunkara, Sundararajan Srinivasan, and Duen Horng Chau, "Best of both worlds: Robust accented speech recognition with adversarial transfer learning," *arXiv preprint arXiv:2103.05834*, 2021.
- [6] Hu Hu, Xuesong Yang, Zeynab Raeesy, Jinxi Guo, Gokce Keskin, Harish Arsikere, Ariya Rastrow, Andreas Stolcke, and Roland Maas, "Redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling," *ICASSP*, 2020.
- [7] Dominika Woszczyk, Stavros Petridis, and David E. Millard, "Domain adversarial neural networks for dysarthric speech recognition," *INTERSPEECH*, 2020.
- [8] Kuan Po Huang, Yu-Kuan Fu, Yu Zhang, and Hung yi Lee, "Improving Distortion Robustness of Selfsupervised Speech Processing Tasks with Domain Adaptation," in *INTERSPEECH*, 2022, pp. 2193–2197.
- [9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.
- [10] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV*. Springer, 2016, pp. 443–450.
- [11] Hang-Rui Hu, Yan Song, Li-Rong Dai, Ian Mcloughlin, and Lin Liu, "Class-aware distribution alignment based unsupervised domain adaptation for speaker verification," *INTERSPEECH*, 2022.
- [12] Wenxin Hou, Jindong Wang, Xu Tan, Tao Qin, and Takahiro Shinozaki, "Cross-domain speech recognition with unsupervised character-level distribution matching," *INTERSPEECH*, 2021.

- [13] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *ICML*, 2019, pp. 1081–1090.
- [14] Jianfei Yang, Han Zou, Yuxun Zhou, Zhaoyang Zeng, and Lihua Xie, "Mind the discriminability: Asymmetric adversarial domain adaptation," in ECCV, 2020.
- [15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.
- [16] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li, "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," *arXiv preprint arXiv:2010.13991*, 2020.
- [17] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao, "Category contrast for unsupervised domain adaptation in visual tasks," in *CVPR*, June 2022, pp. 1203–1214.
- [18] Ankit Singh, "Clda: Contrastive learning for semisupervised domain adaptation," in *NeurIPS*, 2021, vol. 34, pp. 5089–5101.
- [19] Akhil Mathur, Fahim Kawsar, Nadia Berthouze, and Nicholas D. Lane, "Libri-adapt: a new speech dataset for unsupervised domain adaptation," *ICASSP*, 2020.
- [20] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," *INTERSPEECH*, 2021.
- [21] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.
- [22] Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Nadia Berthouze, and Nicholas D. Lane, "Mic2mic: Using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems," in *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, 2019, IPSN '19, p. 169–180.
- [23] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," arXiv preprint arXiv:2007.00991, 2020.