Interspeech 2022
18-22 September 2022, Incheon, Korea

Listen, Adapt, Better WER: Source-free Single-utterance Test-time Adaptation for Automatic Speech Recognition

Guan-Ting Lin[1], *Shang-Wen Li*[2*], *Hung-yi Lee*[1]

[1]National Taiwan University, Taiwan
[2]Amazon AI, USA

r10942104@ntu.edu.tw, hungyilee@ntu.edu.tw

Abstract

Although deep learning-based end-to-end Automatic Speech Recognition (ASR) has shown remarkable performance in recent years, it suffers severe performance regression on test samples drawn from different data distributions. Test-time Adaptation (TTA), previously explored in the computer vision area, aims to adapt the model trained on source domains to yield better predictions for test samples, often out-of-domain, without accessing the source data. Here, we propose the **S**ingle-**U**tterance **T**est-time **A**daptation (**SUTA**[1]) framework for ASR, which is the first TTA study on ASR to our best knowledge. The single-utterance TTA is a more realistic setting that does not assume test data are sampled from identical distribution and does not delay on-demand inference due to pre-collection for the batch of adaptation data. SUTA consists of unsupervised objectives with an efficient adaptation strategy. Empirical results demonstrate that SUTA effectively improves the performance of the source ASR model evaluated on multiple out-of-domain target corpora and in-domain test samples.

Index Terms: Test-time Adaptation, Domain Shift, Speech Recognition

1. Introduction

Deep Learning-based Automatic Speech Recognition (ASR) models achieve impressive success, especially when samples are drawn under the independent and identical distribution (i.i.d.) assumption. However, performance degrades severely when covariate shift (i.e., distribution of test data differs from training data) happens. In real-world scenarios, such covariate-shifted test samples are ubiquitous, making ASR services unstable and unreliable. Therefore, it is critical to alleviate the adverse effect of data shifting.

Unsupervised Domain Adaptation (UDA) is a commonly-used approach that adapts the source model to the target domain without annotated data. The existing UDA approaches such as domain adversarial training [1, 2], knowledge distillation [3], and self-training [4] have shown effectiveness for mitigating data shifting and improving ASR performance. However, they all require access to source data and sufficient target domain examples for adaptation. Such requirement imposes three main *limitation* on the real-world application of UDA: (1) source data is not always available during adaptation due to the privacy/storage issues, (2) latency due to target data pre-collection and heavy computation for model adaptation, and (3) the assumption that the target domain examples come from the same distribution.

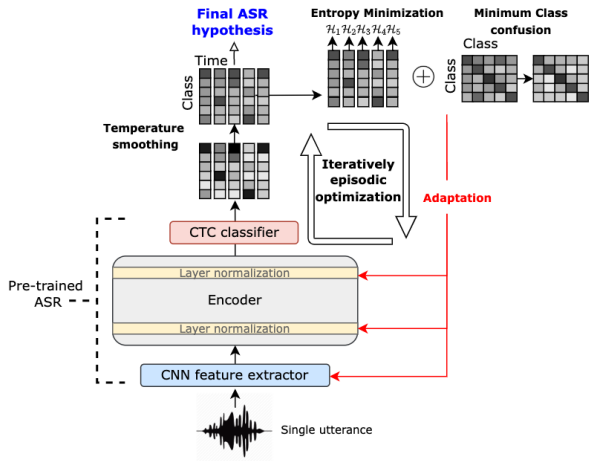

Figure 1: *The illustration of the proposed SUTA framework. Given a single utterance, We leverage Entropy Minimization and Minimum Class Confusion as the unsupervised guidance to adapt the source ASR model in the inference time.*

Test-Time Adaptation (TTA) [5, 6, 7, 8, 9, 10, 11] has recently attracted growing interest since it effectively adapts models in prediction time with little target data (a batch or even a single instance) without access to source data. Several studies have shown remarkable TTA effectiveness in computer vision but lack research attempts on TTA for ASR. It is worth noting that TTA in computer vision is heavily targeted on the Batch Normalization (BN) layer's adaptation by re-estimating batch statistics on target data. However, sequential models, such as ASR models, typically are not equipped with BN layers because the lengths of batched input sequences are different. Therefore, most ASR models use instance-wise Layer Normalization (LN) layer. The discrepancy in data format and model architecture also motivates us to innovate TTA tailored for ASR.

Among different TTA variants, most methods are restricted to batch-level TTA. In other words, TTA resolves the first limitation (i.e., privacy/storage) of UDA, while the rest two (latency and target data distribution) still cause challenges. Recently, SITA [8] has presented a ***single-instance TTA*** method, which lifts the limitation about latency and distribution since it does not require the pre-collecting batch of test samples, and it allows test samples that come from heterogeneous sources. However, SITA still focuses on computer vision problems and relies on data augmentation and statistical estimation of BN layers.

To the best of our knowledge, no existing work attempts TTA or even single-instance TTA on ASR, and there is lacking research exploring the potential of TTA in improving ASR per-

*Work done while working at Amazon Inc. The current affiliation is Meta AI.

[1]https://github.com/DanielLin94144/Test-time-adaptation-ASR-SUTA

10.21437/Interspeech.2022-600

formance. To fill the gap, in this work, we propose the **S**ingle-**U**tterance **T**est-time **A**daptation (**SUTA**) framework to improve ASR test-time performance with a single utterance. SUTA can be applied to any CTC-based end-to-end ASR model. Our method does not rely on large batch size and access to source data, which causes delayed inference and privacy issues; instead, only one testing utterance is needed for test-time adaptation in an unsupervised manner. SUTA consistently improves the source ASR model in multiple out-of-domain target corpora and in-domain test samples with little computational delay.

2. Method

2.1. Problem Formulation

We denote a trained ASR model as $g(y \mid x; \theta)$, parameterized by the pre-trained weight θ, taking utterance x as input and outputs un-normalized score of vocabulary class y. θ can be split to two parts, $\theta_{\mathbf{f}}$ and $\theta_{\mathbf{a}}$. $\theta_{\mathbf{f}}$ is the parameters frozen at adaptation, while $\theta_{\mathbf{a}}$ is updated during the *inference* time. Test corpus D_{test} contains n utterances $\{x_1, x_2, ..., x_n\}$. We focus on adapting θ_a with a single utterance x_i itself in an unsupervised manner. This paper uses a transformer encoder with Connectionist Temporal Classification (CTC) loss [12] as an ASR model, but the proposed approach is model agnostic. In the following, we represent the number of character classes plus one CTC blank token as C, and the number of time frames from the CTC classifier's output as L. The model output is $\mathbf{O} \in \mathbb{R}^{L \times C}$.

2.2. Single-Utterance Test-time Adaptation (SUTA)

We introduce each component of SUTA in the below sections. Figure 1 illustrates the SUTA framework.

2.2.1. Entropy minimization

Since labels are unavailable during testing, we leverage an unsupervised, entropy-based loss function for adaptation. Entropy Minimization (EM) intends to sharpen class distribution, which is a wide-used approach in domain adaptation and semi-supervised learning with a large amount of target data [13, 14, 15, 16, 17, 18], but not TTA. TENT [5] first proposes TTA by EM and adapting BN layer's parameters only with batched inputs. Inspired by TENT, we utilize the EM objective for unsupervised test-time adaptation on model parameters with a single utterance. Since CTC blank token dominates the class distribution in L frames, we exclude those frames where the CTC blank token yields the highest probability to mitigate the class-imbalanced issue. $\mathcal{L}_{em}$ can be calculated as

$$\mathcal{L}_{em} = \frac{1}{L}\sum_{i=1}^{L}\mathcal{H}_i = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{C}\mathbf{P_{ij}}\log\mathbf{P_{ij}}, \tag{1}$$

where $\mathcal{H}_i$ is the entropy at i-th frame, and $\mathbf{P_{ij}}$ is the probability of the j-th class in the i-th frame.

2.2.2. Minimum class confusion

In addition to the EM objective, Minimum Class Confusion (MCC) objective is an alternative adjusting model parameters by reducing the correlation between different classes. This objective is adopted by [19, 20, 21] for domain adaptation, and [10] for batched test-time adaptation in computer vision. The equation of MCC loss is

$$\mathcal{L}_{mcc} = \sum_{j=1}^{C}\sum_{j'\neq j}^{C}\mathbf{P}_{\cdot\mathbf{j}}^{\top}\mathbf{P}_{\cdot\mathbf{j}'}, \tag{2}$$

Algorithm 1 SUTA Algorithm

C : number of classes, L : number of frames, T : temperature, $\theta_{\mathbf{f}}$: frozen weight, $\theta_{\mathbf{a}}$: adaptation weight, x : an utterance in test set D_{test}
for $t = 1$ to N **do**
  $\mathbf{O} = g(y \mid x; \theta_{\mathbf{f}}, \theta_{\mathbf{a}})$
  $\mathbf{P}_{\cdot\mathbf{j}} = \frac{\exp(\mathbf{O}_{\cdot\mathbf{j}}/T)}{\sum_{j'=1}^{C}\exp(\mathbf{O}_{\cdot\mathbf{j}'}/T)}, \forall j \in C$
  $\mathcal{L}_{em} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{C}\mathbf{P_{ij}}\log\mathbf{P_{ij}}$
  $\mathcal{L}_{mcc} = \sum_{j=1}^{C}\sum_{j'\neq j}^{C}\mathbf{P}_{\cdot\mathbf{j}}^{\top}\mathbf{P}_{\cdot\mathbf{j}'}$
  $\mathcal{L} = \alpha\mathcal{L}_{em} + (1-\alpha)\mathcal{L}_{mcc}$
  $\theta_{\mathbf{a}}^{\mathbf{t+1}} = \text{Optimizer}(\theta_{\mathbf{a}}^{\mathbf{t}}, \mathcal{L})$
Output: Decode $g(y \mid x; \theta_{\mathbf{f}}, \theta_{\mathbf{a}}^{\mathbf{N}})$.

where $\mathbf{P}_{\cdot\mathbf{j}} \in \mathbb{R}^{L}$ denotes the probabilities of the j-th class of the L frames. We minimize the class correlation between pairs of class j and j' ($j \neq j'$) by only penalizing non-diagonal values on the class confusion matrix. Actually, [20] showed that the MCC objective is similar to EM, but the gradient is different.

2.2.3. Temperature smoothing

However, we discovered that naively TTA with $\mathcal{L}_{em}$ and $\mathcal{L}_{mcc}$ yield minor performance improvement empirically (Table 2 $T = 1$ row). These results can be explained since entropy loss is small for the confident predictions, so the EM objective cannot gain guidance from those confident frames. Instead, the $\mathcal{L}_{em}$ mainly attributes to the uncertain frames, which may provide unreliable update direction. Very recently, [7] also unrevealed gradient vanishing problem for high confident predictions. To cope with this issue, we use the temperature scaling method to smooth probability distribution, keeping the influence of high-confident frames. The temperature smoothing can also alleviate the negative effects of over-confident prediction for MCC objectives. We smooth the output distribution as

$$\mathbf{P}_{\cdot\mathbf{j}} = \frac{\exp(\mathbf{O}_{\cdot\mathbf{j}}/T)}{\sum_{j'=1}^{C}\exp(\mathbf{O}_{\cdot\mathbf{j}'}/T)}, \tag{3}$$

where $\mathbf{O}_{\cdot\mathbf{j}}$ is the output logits of j-th class for all a time frame. T is larger than 1 for flattening the probability distribution. At every iteration, the smoothed distribution is used for loss calculation in equation (1) and (2).

2.2.4. Training objective

To prevent models from overfitting on one unsupervised objective, both $\mathcal{L}_{em}$ and $\mathcal{L}_{mcc}$ losses are optimized for adaptation. α is the hyper-parameter for weighted-sum two loss. Overall loss function can be written as below:

$$\mathcal{L} = \alpha\mathcal{L}_{em} + (1-\alpha)\mathcal{L}_{mcc} \tag{4}$$

The overview of SUTA is in Algorithm 1. For each utterance x, starting from the original weights of source ASR model, we forward x to obtain model output $\mathbf{O}$, temperature smoothing the distribution, and train the adaptation parameter $\theta_{\mathbf{a}}$ by minimizing the loss function $\mathcal{L}$. After N iterations, the adapted ASR model $g(y \mid x; \theta_{\mathbf{f}}, \theta_{\mathbf{a}}^{\mathbf{N}})$ is used for final inference.

Table 1: *Word Error Rate (%) for different corpora and methods. All WERs are measured without decoding with Language models. State-of-the-art (SOTA) performances are from [22] for LS, [23] for CH, [24] for TD, and [25] for CV. The dashed line "-" means there is no performance reported by prior works.*

Performance reference for source ASR model *wo/ adaptation*	LS test-o + δ 0	0.005	0.01	CH	CV	TD
SOTA (trained on target dataset)	2.5	-	-	5.8	15.4	5.6
RASR [26] (trained on LS)	6.8	-	-	-	29.9	13.0
TTA method						
(1) Our source ASR model [27] (trained on LS *wo/ adaptation*)	8.6	13.9	24.4	31.2	36.8	13.2
(1) + SDPL	8.3	13.1	23.1	30.4	36.3	12.8
(1) + SUTA	**7.3**	**10.9**	**16.7**	**25.0**	**31.2**	**11.9**

3. Experiments

3.1. Source ASR model

We use open-sourced CTC-based ASR model, Wav2vec 2.0-base CTC model [27], as our source ASR model. This model is fine-tuned on Librispeech 960-hour[2] and obtains 3.4 and 8.6 Word Error Rate (WER) on Librispeech test-clean and test-other set. Since this model is trained on Librispeech, we regard Librispeech as the source domain. The architecture of the Wav2vec 2.0-base CTC ASR model is composed of three parts: a convolutional neural network-based feature extractor, a 12-layer transformer encoder, and a linear CTC classifier.

3.2. Datasets

To evaluate the adaptation capabilities of the proposed SUTA approach, we examine the test-time performance in the following target domains: **(1) Librispeech (LS)** [28] contains read speech from audiobook in 16kHz. Although Librispeech test data is in-domain for our ASR source model, we want to verify whether SUTA retains performance for in-domain test samples. Furthermore, we inject additive Gaussian noises for different amplitudes ($\delta = \{0.005, 0.01\}$) on the test-other set to create covariate shift, testing SUTA's capability to mitigate Gaussian noises. **(2) CHiME-3 (CH)** [29] is a noisy version of WSJ corpus with artificial and real-world environmental noises at 16kHz. We utilize the official enhanced evaluation set `et05` to examine SUTA's robustness in noisy acoustic conditions. **(3) Common voice (CV)** [30] is a crowdsourcing project that is supported by volunteers to read Wikipedia sentences and record samples at 48kHz. We re-sample the sampling rate to 16kHz to match the training condition of the source ASR model. Test set from the `En-June-22nd-2020` release version is utilized. **(4) TEDLIUM-v3 (TD)** [31] consists of oratory speech based on TED conference videos. The audio quality is clean and stored at 16kHz. We use the official test set for experiments. For all these datasets, transcripts are pre-processed by uppercasing letters and removing punctuation except for apostrophes following [26].

3.3. Baseline TTA methods

Since there is no existing single-utterance TTA study on ASR, here we propose a naive pseudo labeling approach named Single-utterance dynamic pseudo labeling (**SDPL**) as the baseline method, which uses the source ASR model to predict the pseudo label of the utterance and adapt the model by minimizing CTC loss. The pseudo label is generated by greedy decoding on CTC output and refined at each iteration dynamically. SDPL does not consider the uncertainty of the distribution due to the nature of greedy decoding.

[2]Checkpoint of the trained ASR model is open-sourced in https://huggingface.co/facebook/wav2vec2-base-960h

3.4. Implementation details

The hyperparameters are chosen from LS test-o with noises ($\epsilon = 0.01$), and we use the best-found setup for all other target corpora. We explore different settings of trainable adaptation parameters $\theta_{\mathbf{a}}$ in SUTA. For simplicity, we denote the parameters of layer norm transformation as `LN`, feature extractor as `Feat`, and entire ASR model as `All`. The optimizer is AdamW, and the learning rate is searched from 10^{-4} to 10^{-6}. The best-found learning rate is different for each setup of trainable parameters. Specifically, the best learning rate is $2e^{-4}$, $2e^{-5}$, and $1e^{-6}$ for `LN`, `LN+Feat`, and `All`. It is intuitive as the more parameters are relaxed for adaptation, the lower the learning rate is. We noted that SDPL could only improve WER when adapting with `LN`; more parameters cause drastic WER deterioration (probably due to errors from greedy pseudo label generation). Thus, we only report results of LN for SDPL in Section 3.5.

For default setup in all SUTA experiments, α is 0.3, the number of adaptation iterations *iter* is 10, and $\theta_{\mathbf{a}}$ is `LN+Feat`. Experiments are run on Nvidia Geforce RTX 3090 GPU. Adaptation speed is about 0.115 second for a 1 second utterance for 10-steps adaptation.

3.5. Results

Table 1 summarizes our experiment results. We list the state-of-the-art performances of models trained and evaluated on each dataset as the topline WER in the first row. We then show the RASR's [26] results as a baseline where source (training) and target (evaluation) domains mismatch. RASR is a strong baseline, which leverages a much bigger network (36 transformer layers) and is trained with data augmentation (SpecAugment [32]). We also present results for our source ASR model, Wav2vec 2.0-based CTC model trained with source domain data (LS). Lastly, we show the performance of two single-utterance adaptation techniques, SDPL and our proposed SUTA, on top of our source model. SOTA results are unsurprisingly better than others since the model is trained and evaluated without mismatch. SUTA outperforms our source model and the SDPL baseline consistently. Moreover, our proposed approach yields comparable results to RASR with smaller networks and a training/adaption process, suggesting SUTA's efficacy. Noting that we can also adopt SUTA on RASR's models to improve test-time performance, here the results of RASR are just for showing a strong baseline without adaptation.

Table 2: *Ablation study on CH evaluation set. Besides WER, we report the relative WER reductions (WERR) from Baseline to different ablation settings.*

Specification	WER (%)	WERR (%)
Baseline (unadapted)	31.2	0.0
SUTA best config.		
α=0.3, LN+Feat, T=2.5	25.0	19.9
Weighted loss coefficient (α)		
α=1.0	25.5	18.3
α=0.7	25.3	18.9
α=0.5	25.1	19.6
α=0.0	25.4	18.6
Temperature smoothing (T)		
T=1.0	29.9	4.2
T=1.5	26.7	14.2
T=2.0	25.6	17.9
T=3.0	25.5	18.3
Trainable parameters ($\theta_{\mathbf{a}}$)		
LN	28.8	7.7
Feat	25.1	19.6
All	26.7	14.4

We dive deep into the results. For the in-domain evaluation (i.e., LS without additive noise), we observe that the pseudo labeling baseline (SDPL) slightly improves WER on top of the source ASR model (c.f., 8.3 vs. 8.6). SUTA further improves the WER to 7.3, which is surprising that TTA can even enhance the test-time performance for in-domain samples. With the target domain drifting from source by adding Gaussian noises, WER degrades from 8.6 to 13.9 ($\delta = 0.005$) and 24.4 ($\delta = 0.01$) for our source model. SDPL slightly improves WER, while SUTA drastically reduces WER for 3.9 ($\delta = 0.005$) and 7.7 ($\delta = 0.01$), suggesting that SUTA can improve the robustness of ASR when Gaussian noises are present.

Besides synthesizing the drifted target domain with LS, SUTA mitigates mismatch from multiple real-world audio corpora. For noisy speech in CH, SUTA successfully boosts WER from 31.2 to 25.0, whereas SDPL can only reduce WER to 30.4. Similar results can be found with the crowdsourcing recording in CV and the clean oratory speech in TD. SUTA improves WER from 36.8 to 31.2 for CV and from 13.2 to 11.9 for TD, outperforming SDPL significantly.

3.6. Discussion and analysis

We investigate the importance of different components in SUTA, including the weighted sum coefficient α of the loss function, temperature values, and trainable parameters. Due to space limitations, we only present an analysis on CH.

Ablation study: Ablation studies are summarized in Table 2. Both EM and MCC objectives improve the performance individually, according to row $\alpha = 1.0$ and $\alpha = 0.0$, and weighted-sum combination further enhances relative WER reduction (WERR) from 18.3 / 18.6 to 19.9 when $\alpha = 0.3$. The results indicate that using multiple objectives can improve the TTA, preventing overfitting on a single loss. Besides, the contribution of temperature smoothing is significant, reducing from 29.9 (T = 1.0) to 25.0 (T= 2.5) WER.

Choosing appropriate trainable parameters is also crucial to SUTA. There is 7.7 WERR by solely adapting the layer normalization (`LN`). If we fine-tune the normalization layer and the feature extractor (`LN+Feat`), we can further increase performance gain to 19.9 WERR. Nevertheless, relaxing too many parameters for single-utterance fine-tuning causes slight performance degradation, as indicated by the result for `All`, which only yields 14.4 WERR.

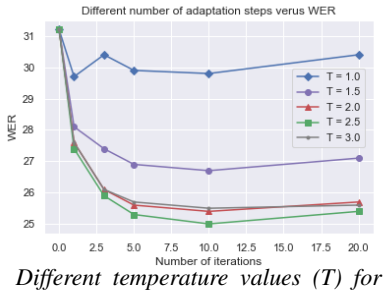

Figure 2: *Different temperature values (T) for temperature smoothing at different number of iterations on CHiME-3 evaluation set.*

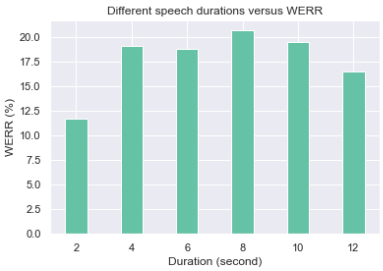

Figure 3: *The analysis of how utterance duration influence the SUTA effectiveness on CHiME-3 evaluation set.*

Iteration step: We analyze the different number of model adaptation steps and their impact on performance. We explore 1, 3, 5, 10, and 20 for the number of iterations. Figure 2 demonstrates that the WER is consistently reduced before 10 steps; however, after 10 steps, the improvement saturates, particularly when not adopting temperature smoothing ($T = 1$). On the other hand, utilizing temperature smoothing ($T > 1$) stabilizes adaptation, mitigates the performance degradation in the later iterations, and enhances overall effectiveness.

Utterance length: When the length of input audio is short, it only contains one or two words. In this case, the output prediction merely includes a small number of classes and may be prone to class collapse when entropy is minimized. We investigate the relationship between utterance length and WERR in Figure 3. The result indicates that the WERR of short utterances (less than 2 seconds) is about 12 % WERR, which is much lower than the around 17.5 % WERR observed in typical utterances (those longer than 2 seconds). Although the performance gain of short audio is smaller than the normal one, we do not observe any adverse effect, and there is still WER improvement.

4. Conclusion

In this work, we propose the first source-free **S**ingle-**U**tterance **T**est-time **A**daptation (SUTA) framework on ASR, which can efficiently adapt CTC-based ASR models given one target utterance in the inference time. Specifically, we use entropy minimization and minimum class confusion objectives for TTA and further improve TTA by temperature smoothing. Experimental results demonstrate that SUTA effectively reduces the source ASR model's WER on several out-of-domain corpora, even enhancing the performance of in-domain test samples. We plan to design TTA on sequence-to-sequence ASR models in the future, improving performance on short utterances.

5. References

[1] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[2] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[3] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," *Proc. Interspeech 2017*, pp. 2386–2390, 2017.

[4] 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)*. IEEE, 2021, pp. 6553–6557.

[5] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2020.

[6] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6028–6039.

[7] C. K. Mummadi, R. Hutmacher, K. Rambach, E. Levinkov, T. Brox, and J. H. Metzen, "Test-time adaptation to distribution shift by confidence maximization and input transformation," *arXiv preprint arXiv:2106.14999*, 2021.

[8] A. Khurana, S. Paul, P. Rai, S. Biswas, and G. Aggarwal, "Sita: Single image test-time adaptation," *arXiv preprint arXiv:2112.02355*, 2021.

[9] X. Hu, M. G. Uzunbas, S. Chen, R. Wang, A. Shah, R. Nevatia, and S.-N. Lim, "Mixnorm: Test-time adaptation through online normalization estimation," *ArXiv*, vol. abs/2110.11478, 2021.

[10] F. You, J. Li, and Z. Zhao, "Test-time batch statistics calibration for covariate shift," *arXiv preprint arXiv:2110.04065*, 2021.

[11] F. Fleuret *et al.*, "Test time adaptation through perturbation robustness," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[13] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulo, "Autodial: Automatic domain alignment layers," in *2017 IEEE international conference on computer vision (ICCV)*. IEEE, 2017, pp. 5077–5085.

[14] R. Shu, H. H. Bui, H. Narui, and S. Ermon, "A dirt-t approach to unsupervised domain adaptation," in *ICLR (Poster)*, 2018.

[15] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9471–9480.

[16] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[17] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8050–8058.

[18] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in *European Conference on Computer Vision*. Springer, 2020, pp. 464–480.

[20] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2090–2099.

[21] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[22] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *arXiv preprint arXiv:2108.06209*, 2021.

[23] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[24] W. Zhou, W. Michel, K. Irie, M. Kitza, R. Schlüter, and H. Ney, "The rwth asr system for ted-lium release 2: Improving hybrid hmm with specaugment," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7839–7843.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[26] 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.

[27] 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. 12 449–12 460, 2020.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[30] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, May 2020.

[31] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *International conference on speech and computer*. Springer, 2018, pp. 198–208.

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.