Robust Automatic Speech Recognition via WavAugment Guided Phoneme Adversarial Training

Gege Qi[1], *Yuefeng Chen*[1], *Xiaofeng Mao*[1], *Xiaojun Jia*[2], *Ranjie Duan*[1], *Rong Zhang*[1], *Hui Xue*[1]

[1]Alibaba Group, China
[2]Chinese Academy of Sciences, China
{qigege.qgg,yuefeng.chenyf,mxf164419}@alibaba-inc.com

Abstract

Developing a practically-robust automatic speech recognition (ASR) is challenging since the model should not only maintain the original performance on clean samples, but also achieve consistent efficacy under small volume perturbations and large domain shifts. To address this problem, we propose a novel WavAugment Guided Phoneme Adversarial Training (WAPAT). WAPAT use adversarial examples in phoneme space as augmentation to make the model invariant to minor fluctuations in phoneme representation and preserve the performance on clean samples. In addition, WAPAT utilizes the phoneme representation of augmented samples to guide the generation of adversaries, which helps to find more stable and diverse gradient-directions, resulting in improved generalization. Extensive experiments demonstrate the effectiveness of WAPAT on End-to-end Speech Challenge Benchmark (ESB). Notably, SpeechLM-WAPAT outperforms the original model by 6.28% WER reduction on ESB, achieving the new state-of-the-art.

Index Terms: robust automatic speech recognition, data augmentation, adversarial training

1. Introduction

Nowadays, there have been remarkable advancements in Deep Neural Network (DNN) based Automatic Speech Recognition (ASR) [1], resulting in the emergence of numerous speech-related applications that assist humans in their daily activities. However, despite the impressive performance of ASR systems, they are limited to specific tasks since they assume that the training and testing data are drawn from the same distribution [2]. Thus, applying ASR in real-world applications under diverse environment is still a huge challenge [3, 4, 5].

In this work, we aim to address such a challenging cross-domain scenario where an ASR system needs to be robust against various potential distortion. However, there are two major challenges: 1) **Robustness against perturbation:** Real-world volume perturbation (*e.g.*, environmental noise, reverberation, and background speakers) significantly impacts the performance of an ASR system [6, 7]. 2) **Robustness Generalization:** There exist various type of volume perturbation in practical scenario. However, a ASR is robust against one type of perturbation not promised being robust under unknown domain (*e.g.*, change of speaking style). Existing work either adopt data augmentation to improve ASR's robustness against specific perturbation but limited under unseen domain [8, 9, 10, 11], or use speech enhancement as a pre-processing to deal with various potential noise [12]. Both of them fail to achieve a real-robust ASR system which can be applied in the real world. Therefore, enhancing the robustness of ASRs while improving their robustness generalization across different perturbation remains a significant challenge.

To address the challenge, we propose a novel method called **W**av**A**ugment Guided **P**honeme **A**dversarial **T**raining (WAPAT) by leveraging adversarial training (AT) technique. Though previous work [13] claimed AT results in a trade-off between robustness and clean accuracy. However, research in the field of natural language processing [14] and recent in computer vision [15] demonstrated that aligning the distributions of adversarial and original samples during AT can benefit robustness and clean accuracy simultaneously. Borrowing the idea, we propose applying AT on phoneme space to create adversarial speech with realistic semantic. To be specific, WAPAT employs a single-step attack to generate adversarial perturbations on phoneme representations. A data augmentation is further applied to guide the attack for generating more stable and diverse phoneme adversarial examples. In detail, with the time-domain WavAugment [16] technique, we use Kullback-Leibler Divergence (KLD) to align the adversaries on original samples and those on augmented samples. Furthermore, multiple augmentations in WavAugment are used to guide the adversarial training for searching for diverse gradient directions and leading to better generalization. Figure 1 shows the overall pipeline of WavAugment guided phoneme adversarial training.

We further explore the impact of individual techniques in WavAugment and their combinations with WAPAT on the performance of the model. Our findings indicate that while hard augmentations can improve robustness on some datasets, they fails on others. It is reasonable to expect challenges in generalizing audio augmentations across different domains, given the inherent complexity of audio signals. Instead, our WAPAT consistently improves performance in terms of both cross-domain datasets and different types of transformations. The stability of generalization indicates that the WavAugment-guided adversary is effective in inducing robust features into target ASR.

In summary, we make the following contributions:

- To our knowledge, this is the first work that sheds light on adversarial training on phoneme-unit space for improving standard performance and generalization simultaneously.
- We propose WavAugment Guided Phoneme Adversarial Training (WAPAT), which employs phoneme representation of the augmented audios to guide the generation of adversaries, resulting in more diverse robust features.
- By combining SpeechLM [17] pre-training and WAPAT finetuning, our method achieves new state-of-the-art performance on the challenging benchmark ESB [18], which contains multiple speech datasets from a broad range of domains.

10.21437/Interspeech.2023-1556

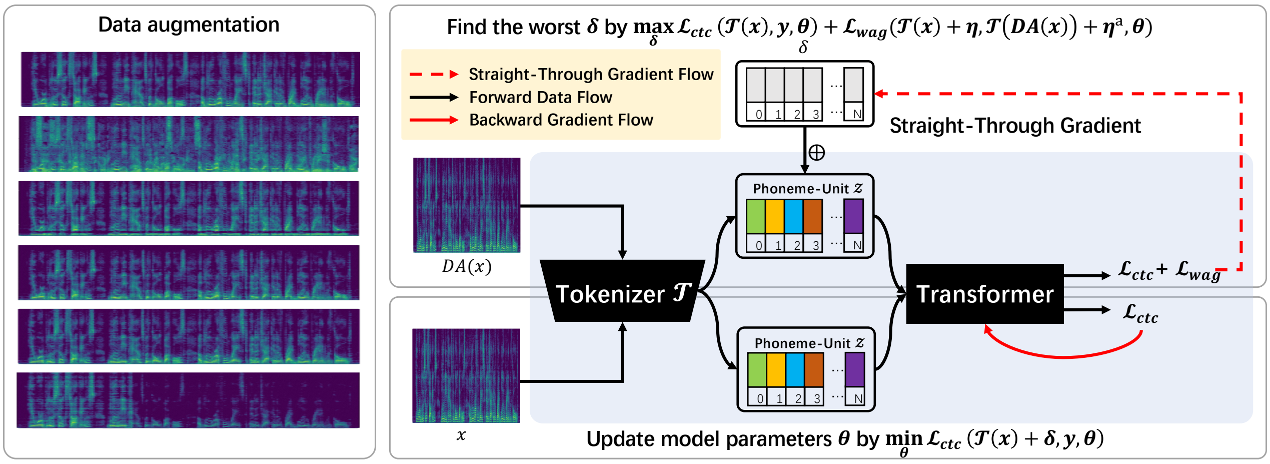

Data augmentation
Find the worst δ by $\max_{\delta} \mathcal{L}_{ctc}(\mathcal{T}(x), y, \theta) + \mathcal{L}_{wag}(\mathcal{T}(x) + \eta, \mathcal{T}(DA(x)) + \eta^{a}, \theta)$
Straight-Through Gradient Flow
Forward Data Flow
Backward Gradient Flow
δ
0 1 2 3 … N
Straight-Through Gradient
$\oplus$
Phoneme-Unit $\mathcal{Z}$
0 1 2 3 … N
$DA(x)$
Tokenizer $\mathcal{T}$
Transformer
$\mathcal{L}_{ctc} + \mathcal{L}_{wag}$
$\mathcal{L}_{ctc}$
Phoneme-Unit $\mathcal{Z}$
0 1 2 3 … N
x
Update model parameters θ by $\min_{\theta} \mathcal{L}_{ctc}(\mathcal{T}(x) + \delta, y, \theta)$

Figure 1: *The overview of our proposed WavAugment Guided Phoneme Adversarial Training (*WAPAT*). Left: from top to bottom, the figures depict the log mel spectrogram of the base input with no augmentation, additive noise, band reject, pitch modification, time masking and reverberation applied. Right: the pipeline of* WAPAT*, where the augmented samples are used to guide the generation of adversaries during adversarial training on the symbolic phoneme space.*

2. Related Work

Existing approaches for improving ASR generally from two aspects: **1) Improving ASR's robustness against specific perturbation:** Early works have shown that several data augmentation methods, such as vocal tract length perturbation [8], volume perturbation [9] and speed perturbation [11], can improve the robustness of ASR models. SpecAugment [10] is widely used to train ASR models due to its efficiency. Specifically, SpecAugment randomly masks chunks of time or frequency channels on spectrograms. However, these DA techniques are typically designed manually for specific domains based on domain-specific knowledge and experience. When dealing with an unknown target domain or multiple domains, it can be challenging for experts to apply specific transformations, or to construct and fine-tune more sophisticated augmentation compositions [19]. Besides data augmentation, several work utilize adversarial training [20, 21] aiming to improve ASR's adversarial robustness under adversarial examples. However, all of these work target ASR's robustness under specific perturbation, and the improved ASR is still limited on unseen domain. **2) Improving ASR's performance via pre-processing:** To achive better performance, there are also several works propose using speech enhancement methods to remove noise from speech signals before passing them through a standard ASR [12]. A more recent approach that operates on raw waveform for real-time speech enhancement is [22]. However, these methods often rely on front-end processing modules, which decrease efficiency and add computational overhead. Also, the speech enhancement method do not really improve the robustness of ASR itself. In this paper, we aim to build a truly robust ASR which is robust under multiple or even unseen perturbations. It has the potential to be applied in various applications in real-world setting.

3. Method

3.1. Adversarial Training on ASR

Consider the training utterance and text label set $\mathcal{D} = \{(x_i, y_i)_{i=1}^{n}\}$, an ASR model with learnable parameters θ, and a recognition objective given by Connectionist Temporal Classification (CTC) loss $\mathcal{L}_{ctc}$. Adversarial Training (AT) aims to optimize θ by solving a minimax optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta} \mathcal{L}_{ctc}(x + \delta, y, \theta)\right] \; s.t.||\delta||_p \leq \epsilon, \quad (1)$$

where the inner optimization seeks perturbations δ on speech values that maximize the loss, and the outer minimization update θ to improve the worst-case performance of the network. The boundary $||\delta||_p \leq \epsilon$ restricts the magnitude of the perturbation. We use projected gradient descent (PGD) [23] for the inner optimization. In the following section, we will tackle the challenges of mitigating performance degradation and enhancing the generalization ability of ASR models through AT techniques.

3.2. Phoneme Adversarial Training

We borrow the perspective of the AT on the contextualized language representation, and propose a new Phoneme Adversarial Training (PAT) for ASRs, *i.e.*, conducting AT on the phoneme representation space instead of raw input space. To accomplish this, we leverage the SpeechLM framework proposed by [17] to recognize speeches. The phoneme unit sequence of input x can be obtained by applying a transformer based phoneme-unit tokenizer $\mathcal{T}$. In the inner maximization step of AT, we generate phoneme adversarial examples by slightly modifying Equation 1. The objective of PAT can be formulated as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta} \mathcal{L}_{ctc}(\mathcal{T}(x) + \delta, y, \theta)\right] \; s.t.||\delta||_\infty \leq \epsilon, \quad (2)$$

where ϵ is the magnitude of the perturbation, PAT finds the failure case, *i.e.*, $\mathcal{T}(x) + \delta$ in phoneme space.

3.3. WavAugment Guided Phoneme Adversarial Training

For improving the generalization of ASR by AT, we aim to generate adversarial examples that exhibit both stability and diversity. Towards this end, we propose a novel WavAugment Guided Phoneme Adversarial Training (WAPAT) method.

Adversarial examples are typically distributed near the decision boundary, and slight variations can cause them to lose

Algorithm 1: Pseudo code of WAPAT

```
Input: Speech tokenizer $\mathcal{T}$; A sampled mini-batch of clean
       audios $x$ with labels $y$; Perturbation size $\epsilon$.
Output: Learned network parameter $\theta$
Fix the network parameters of $\mathcal{T}$
for each training steps do
    $z \leftarrow \mathcal{T}(x)$
    $z \leftarrow \mathcal{U}(\mathcal{B}_{\epsilon}^{\infty}(z))$    //Initialize adversarial example
    $z^a \leftarrow \mathcal{T}(\mathcal{DA}(x))$
    $\eta, \eta^a \leftarrow \nabla_z \mathcal{L}_{ctc}(z, y, \theta), \nabla_z \mathcal{L}_{ctc}(\mathcal{T}(z^a, y, \theta)$
    $\delta \leftarrow \nabla_z [\mathcal{L}_{ctc}(z, y, \theta) + \mathcal{L}_{wag}(z + \eta, z^a + \eta^a, \theta)]$
    $\hat{z} \leftarrow \prod_{\mathcal{B}_{\epsilon}^{\infty}(z)} (z + \delta)$    //Generate adversarial examples
    Update model parameter on $\mathcal{L}_{ctc}(\hat{z}, y, \theta)$
end for
```

their adversarial nature [24]. Therefore, enhancing the stability of adversarial examples is beneficial for obtaining more robust features. To tackle the instability problem, we introduce the WavAugment guided term along with the CTC loss to form a new objective function during the generation of adversarial examples.

The WavAugment operation, denoted as $\mathcal{DA}(\cdot)$, applies time-domain data augmentation to an audio sample. We represent the phoneme representation of the original sample and the augmented sample as z and z^a, respectively. Then, the perturbations generated for the two samples are denoted as η and η^a. The loss WavAugment guided term encourages the predictions of the adversarial examples of the original and augmented samples to be similar. Formally, the objective function can be written as:

$$\mathcal{L}_{wag}(z+\eta, z^a+\eta^a, \theta) = -\mathcal{D}_{KL}\left[p(z+\eta, \theta) || p(z^a+\eta^a, \theta)\right]$$
$$z = \mathcal{T}(x), z^a = \mathcal{T}(\mathcal{DA}(x)) \quad (3)$$

where $p(x|\theta)$ is the joint probability and $\mathcal{D}_{KL}$ is the KL-divergence. From an optimization perspective, the WavAugment guided term helps in avoiding local optima during the perturbation generation process, leading to the creation of more stable and robust features for ASR models.

In this paper, the basic data augmentations of WavAugment [16] is reserved, including pitch modification (pitch), additive noise (add), band reject filtering (band_rej), time masking (time_mask) and reverberation (reverb). pitch and add are intended to simulate variations in the speaker's voice and environmental noise. band_rej and time_mask augmentations can introduce noise into the neural representation of speech, which can help the model learn to better handle noisy speech. The reverb simulates the effect of sound reflections in a room, which can help the model learn to better handle the effects of reverberation in real-world environments. Here, we use gpuRIR [25] to obtain acoustic room impulse responses.

To enhance the diversity of adversarial examples, we utilize all of the augmentations available in WavAugment to guide the generation process. During training, one of the transformations from WavAugment is applied to each batch of samples. In Figure 1, we show an example of log mel spectrograms augmented with different transformations. Further details regarding the WAPAT can be found in Algorithm 1. Given a SpeechLM based speech recognition model, the speech transformer $\mathcal{T}$ first yields a higher level phoneme representation z from speech input. The WavAugment guided perturbation δ can be obtained by computing the gradient of z towards maximizing the $\mathcal{L}_{ctc}$ and $\mathcal{L}_{wag}$. For clarity, $\mathcal{B}_{\epsilon}^{\infty}(z) := \{z' : ||z' - z||_{\infty} \leq \epsilon\}$ defines a ball of radius ϵ around z in the l_{∞} norm. The symbol $\mathcal{U}$ denotes the uniform distribution, and $\prod$ denotes a projection function. Finally, the adversarial example $\hat{z}$ is fed into models for training.

4. Experiment

Datasets and Settings We conducted experiments on the ESB [18] benchmark to evaluate cross-domain ASR robustness. ESB comprises eight datasets with a broad range of domains, acoustic conditions, speaker styles, and transcription requirements. Notably, Librispeech [26] and Common Voice [27] only contain narrated style speeches, while VoxPopuli [28] and TED-LIUM [29] have oratory style speeches, and AMI [30] contains spontaneous style speeches. GigaSpeech [31], SPGISpeech [32], and Earnings-22 [33] cover two different styles of speeches. Additionally, We included the optional CHiME-4 [34] dataset with narrated style to test generalization. We use the standard split of the above datasets and unify the transcription format as normalised. We finetune the SpeechLM-P model[1] on the Librispeech-100h dataset, which is pre-trained on both the LibriSpeech-960h audio and the LibriSpeechLM corpus2. And we evaluate the robustness on datasets in ESB. Audio format is 16-bit WAV with 16 kHz, and transcription format is unified into the normalized form.

Implementation Details The hyper-parameters used in WavAugment are as follows: pitch randomly modifies the pitch of the waveform by $n \in [-300, 300]$ semitones. add randomly adds noise from MUSAN [35] dataset with a scaled signal-to-noise ratio between $[0, 40]$. The maximal width of the rejected spectrum in band_rej is 150 Hz. The time_mask operation zeros out ten random subsequences of the inputs with a maximum length of 2000 ms. The room dimensions and other parameters in reverb are randomly sampled within default ranges [2].

We evaluate the accuracy of our predictions against target transcriptions using the word error rate (WER). The ESB score is the macro-averaged value of datasets in the ESB benchmark, excluding Librispeech. We implement WAPAT on the pre-trained SpeechLM-P [17], which consists of a Speech Transformer, a Shared Transformer and a CTC head. By default, we refer SpeechLM-P-Base to SpeechLM in all tables and figures. Models are optimized by Adam with a maximum learning rate of 1e-5 and a tri-stage learning rate schedule with the warming-up, holding, and decay periods of [0.1, 0.4, 0.5]. We train the models for a total of 30K steps with a batch size of 800 seconds. Perturbations are bounded with an l_{∞}-norm of 0.01. All experiments are conducted on four NVIDIA Tesla A100.

4.1. Overall Performance

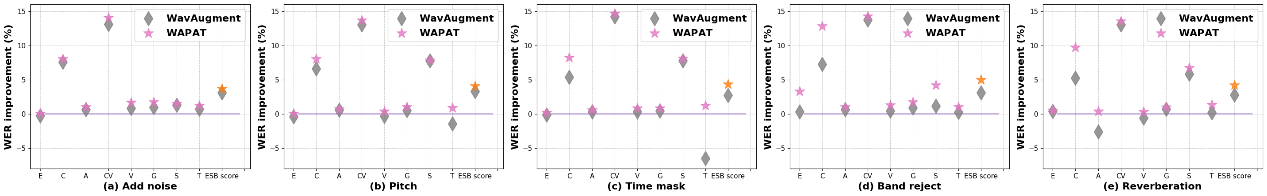
To demonstrate the effectiveness of WAPAT, we first compared it with data augmentation and adversarial training methods. We make a fair comparison with standard WavAugment [10] and SpecAugment [10]. Although SpecAugment performs well on Librispeech test datasets, it shows poor performance in terms of robustness on ESB. In addition, WavAugment has the suboptimal performance of robustness, with an ESB score of 34.18. Notably, our WAPAT achieves superior performance compared to the above data augmentation methods on both in-domain and out-of-domain datasets by a large margin. Compared with the waveform space AT method AdvEx [20], WAPAT achieves 10.01% improvement in ESB score. This further verifies the

[1] https://github.com/microsoft/SpeechT5/tree/main/SpeechLM

[2] https://github.com/DavidDiazGuerra/gpuRIR

Table 1: *WER comparison on the ESB benchmark over various methods for enhancing the robustness of ASR. Best performances are highlighted in bold.*

Method	Librispeech test-clean	Librispeech test-other	Chime-4	Common Voice	VoxPopuli	TED-LIUM	GigaSpeech	SPGISpeech	Earnings-22	AMI	ESB score
SpecAugment [10]	3.32	7.34	45.49	38.46	36.47	19.03	24.57	20.10	51.10	45.02	36.19
WavAugment [16]	3.34	7.35	35.65	38.16	36.64	18.12	25.53	18.99	52.79	46.10	34.18
AdvEx [20]	3.36	7.36	46.10	38.35	36.74	18.18	24.49	19.36	52.01	44.79	36.24
DEMUCS [22]	3.33	7.29	33.57	43.63	36.71	18.31	24.39	26.24	56.76	44.63	35.32
WAPAT	**3.32**	**7.28**	**32.68**	**36.43**	**36.38**	**18.12**	**24.25**	**18.40**	**49.78**	**44.53**	**32.58**



(a) Add noise (b) Pitch (c) Time mask (d) Band reject (e) Reverberation

Figure 2: *Comparison of the WER improvement on ESB benchmark, including Earnings-22 (E), CHiME-4 (C), AMI (A), Common Voice (CV), VoxPopuli (V), GigaSpeech(G), SPGISpeech (S) and TED-LIUM (T) dataset. The last column is the ESB score.*

Table 2: *Ablation study of the proposed* WAPAT *on cross-domain datasets, (a) is different adversarial training variant, (b) is magnitude ϵ.*

Method	Librispeech test-clean	Librispeech test-other	ESB Score
(a) NO-AT	3.34	7.38	36.47
w/ PHONEME AT	3.32	7.34	35.18
w/ WAVAUGMENT PAT	**3.32**	**7.28**	**32.58**
(b) WAPAT			
$\epsilon = 0.005$	3.32	7.35	34.42
$\epsilon = 0.01$	**3.32**	**7.28**	**32.58**
$\epsilon = 0.015$	3.32	7.31	33.24

strengths of our proposed WAPAT in terms of generalization on the phoneme space. Interestingly, WAPAT shows obvious advantages on Chime-4 and Common Voice datasets, which share the same speaking style (Narrated) as the LibriSpeech set. To provide a more comprehensive evaluation, we test the SpeechLM with the speech enhancement-based method DEMUCS [22]. With sacrificing of some computational efficiency, DEMUCS achieves good performance on generalization, however, still inferior to our method.

4.2. Discussion

We further explore the impact of individual techniques in WavAugment and their combinations with WAPAT on the performance of the model, as shown in Figure 2. Specifically, we report the percentage of WER reduction for both standard WavAugment and our WAPAT, compared to the baseline model.

It can be seen that WavAugment is a useful technique for improving the robustness of models. However, there are cases where the individual augmentation perform worse than the baseline on certain datasets. For example, `time_mask` increases the WER on the TED-LIUM dataset. Furthermore, we note that with the same transformation, the WER reduction of WAPAT is greater than that of WavAugment. Additionally, for all transformations, there are some oscillations in WavAugment while WAPAT is consistently increased compared to the baseline. The results accords with the expected that phoneme adversarial training with WavAugment guidance constrains stable optimization of adversaries, resulting in better generalization.

4.3. Ablation Study

As shown in Table 2, to better understand the function of each component of WAPAT, ablation studies are performed and expected to answer the following questions.

How effective is the PAT? Echoing (a) in Table 2, SpeechLM–P with proposed phoneme adversarial training (PHONEME AT) can achieve the better performance on in-domain and out-of-domain datasets than baseline (NO-AT). It indicates adversarially altered phoneme perturbations are much closer to the clean distribution, while strengthen the robustness by capturing more robust features.

Is WAPAT superior than PAT? With WAVAUGMENT PAT means that PAT is guided by the WavAugment, *i.e.*, proposed WAPAT. The ESB score of WAPAT has decreased by roughly 7.4% when compared to PAT. It is evident that WavAugment guidance AT indeed aids in finding stronger robust features.

Does the choice of magnitude ϵ matter? We present the WAPAT results with different magnitude ϵ in Table 2 (b). $\epsilon = 0$ means the standard training of SpeechLM (NO-AT), which makes the models have the worst performance on clean WER and robustness. With the increase of ϵ to 0.01, there is a drop of both clean WER and ESB score. Moreover, we find the clean WER of target model has the lower sensibility on ϵ. But with the ϵ becoming larger, AT greatly damages the generalization, *e.g.*, with $\epsilon = 0.015$, ESB score increases to 33.24. This finding is also revealed by [36].

5. Conclusions and Limitations

In this paper, we propose a novel WavAugment Guided Phoneme Adversarial Training (WAPAT) method, to enhance the cross-domain generalization of ASR systems. WAPAT utilizes the phoneme representation of augmented audios to guide the generation of adversarial examples, resulting in consistently stronger generalization on multiple datasets without sacrificing clean performance. Our experiments demonstrate that WAPAT achieves state-of-the-art robustness on challenging ESB benchmark. However, WAPAT still costs increased training time, this limitation also holds for any adversarial training. This limitation is remained as the future optimization direction.

6. References

[1] 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.

[2] P. P. Parada, A. Dobrowolska, K. Saravanan, and M. Ozay, "pmct: Patched multi-condition training for robust speech recognition," *arXiv preprint arXiv:2207.04949*, 2022.

[3] Y. Hu, N. Hou, C. Chen, and E. S. Chng, "Dual-path style learning for end-to-end noise-robust speech recognition," *arXiv preprint arXiv:2203.14838*, 2022.

[4] R. Fan and A. Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children's asr," *arXiv preprint arXiv:2206.07931*, 2022.

[5] Y. Hu, C. Chen, R. Li, Q. Zhu, and E. S. Chng, "Gradient remedy for multi-task learning in end-to-end noise-robust speech recognition," *arXiv preprint arXiv:2302.11362*, 2023.

[6] B. Gajic and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 600–608, 2006.

[7] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[8] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[9] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[10] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[12] K. Tan and D. Wang, "Improving robustness of deep learning based monaural speech enhancement against processing artifacts," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6914–6918.

[13] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[14] M. Ivgi and J. Berant, "Achieving model robustness through discrete adversarial training," *arXiv preprint arXiv:2104.05062*, 2021.

[15] X. Mao, Y. Chen, R. Duan, Y. Zhu, G. Qi, S. Ye, X. Li, R. Zhang, and H. Xue, "Enhance the visual representation via discrete adversarial training," *arXiv preprint arXiv:2209.07735*, 2022.

[16] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P.-E. Mazaré, M. Douze, and E. Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 215–222.

[17] Z. Zhang, S. Chen, L. Zhou, Y. Wu, S. Ren, S. Liu, Z. Yao, X. Gong, L. Dai, J. Li *et al.*, "Speechlm: Enhanced speech pre-training with unpaired textual data," *arXiv preprint arXiv:2209.15329*, 2022.

[18] S. Gandhi, P. Von Platen, and A. M. Rush, "Esb: A benchmark for multi-domain end-to-end speech recognition," *arXiv preprint arXiv:2210.13352*, 2022.

[19] R. Damania, C. Homan, and E. Prud'hommeaux, "Combining simple but novel data augmentation methods for improving low-resource asr," 2022.

[20] S. Sun, C.-F. Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Training augmentation with adversarial examples for robust speech recognition," *arXiv preprint arXiv:1806.02782*, 2018.

[21] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3107–3111.

[22] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[25] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, pp. 5653–5671, 2021.

[26] 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.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[28] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[29] A. Rousseau, P. Deléglise, and Y. Esteve, "Ted-lium: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.

[30] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[31] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[32] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman *et al.*, "Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv preprint arXiv:2104.02014*, 2021.

[33] M. Del Rio, P. Ha, Q. McNamara, C. Miller, and S. Chandra, "Earnings-22: A practical benchmark for accents in the wild," *arXiv preprint arXiv:2203.15591*, 2022.

[34] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[35] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[36] K. Kireev, M. Andriushchenko, and N. Flammarion, "On the effectiveness of adversarial training against common corruptions," in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 1012–1021.