INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatic Data Augmentation for Domain Adapted Fine-Tuning of Self-Supervised Speech Representations

Salah Zaiem[1], *Titouan Parcollet*[2,3], *Slim Essid*[1]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]Samsung AI Center, Cambridge, United-Kingdom
[3]University of Cambridge, United-Kingdom
salah.zaiem@telecom-paris.fr

Abstract

Self-Supervised Learning (SSL) has allowed leveraging large amounts of unlabeled speech data to improve the performance of speech recognition models even with small annotated datasets. Despite this, speech SSL representations may fail while facing an acoustic mismatch between the pretraining and target datasets. To address this issue, we propose a novel supervised domain adaptation method, designed for cases exhibiting such a mismatch in acoustic domains. It consists in applying properly calibrated data augmentations on a large clean dataset, bringing it closer to the target domain, and using it as part of an initial fine-tuning stage. Augmentations are automatically selected through the minimization of a conditional-dependence estimator, based on the target dataset. The approach is validated during an oracle experiment with controlled distortions and on two amateur-collected low-resource domains, reaching better performances compared to the baselines in both cases.
Index Terms: self-supervised learning, domain adaptation.

1. Introduction

Self-supervised learning (SSL) enables the use of large amounts of unlabelled data to obtain substantial performance improvements in a variety of downstream tasks without relying on manual annotations. Various approaches have been introduced including predictive coding [1, 2], multi-task learning [3, 4], auto-encoding techniques [5] or contrastive learning [6, 7]. In this context, data augmentation has become an important part of many self-supervised approaches. Particularly, various studies have shown that applying several distortions during pretraining leads to more robust representations, either with Contrastive Predictive Coding (CPC) [8], or with Wav2vec2.0 [9, 10]. Recently, WavLM [11] incorporated distortions to add a denoising criterion to its predictive objective.

However, and despite its success, self-supervised learning has been shown to suffer from domain mismatch where the fine-tuning samples from the target domain are vastly different from the pretraining ones [12, 10]. While progress has been made in achieving near-optimal performance on clean datasets such as LibriSpeech, spontaneous speech datasets and non-professionally recorded ones still exhibit lower performance, as displayed in recent speech SSL benchmarks [13, 14].

To mitigate the performance drop caused by domain mismatch, various domain adaptation techniques have been explored, particularly in transfer learning settings [15]. In the self-supervised context, adversarial approaches have been applied during the unsupervised pretraining and tested on speech recognition [16, 17], emotion recognition [18] and speaker recognition [19]. Along with domain adversarial paradigms, Huang *and al.* [20] investigated continual learning methods during pretraining. Distinctly, our method does not aim at aligning latent representations, but rather transforms the audio waveforms of a neutral dataset to match the acoustic conditions of the target domain using data augmentations, rendering this dataset better suited to the final task in an initial fine-tuning stage.

Furthermore, retraining the self-supervised feature extractors with additional domain-invariant enforcement, as proposed in the literature, is a hard and costly endeavor, with the latest SSL models being trained on $94k$ hours of audio data using 64 V100 GPUs [11]. Thus, we envisage the option of augmenting a supposedly neutral dataset and using it for the first fine-tuning step. The augmentations to be applied and their parameters are chosen in order to optimize the similarity in terms of recording conditions between the modified and the target dataset and hence the final performance. Our method presents three main advantages. First, it enables the use of large and clean available annotated datasets, enhancing the textual diversity of the training corpus. Second, it does not require a new pretraining as it directly fine-tunes available SSL models. Finally, it allows an efficient data augmentation exploration, as the selection and parametrization is automatic and does not involve any neural network training. It is, thus, largely more efficient than thorough testing, as scoring 200 augmentation policies takes 3 hours on 10 CPUs, while complete testing of one augmentations distribution necessitates around 20 hours of GPU computations.

The contributions of this work are two-fold: i) Propose a new method for supervised domain adaptation consisting in applying appropriate signal distortions to a clean labeled dataset used for an initial fine-tuning step. The method is validated with an oracle simulated experiment and experiments with naturally noisy datasets. ii) Release the code base, implemented with SpeechBrain [21] for replication and further improvements.[1]

Figure 1 presents an overview of the method, summarizing the three steps conducted for every considered target dataset. First, and given the labeled target dataset, an augmentation distribution is automatically selected (Section 2). Second, a first fine-tuning of the self-supervised representation is done, using the neutral dataset distorted with the augmentations selected in the first step. Finally, a second fine-tuning on the small target domain dataset is done leading to the final model that will be evaluated using the target test set (Section 3.3). Experiments show a speech recognition performance relative improvement reaching 19.5% in a real-world distorted dataset scenario.

2. Selecting the Augmentation Distribution

Given a labeled target speech recognition dataset, our method selects an augmentation distribution that is best suited to its

[1]https://github.com/salah-zaiem/augmentations_adaptation

10.21437/Interspeech.2023-1040

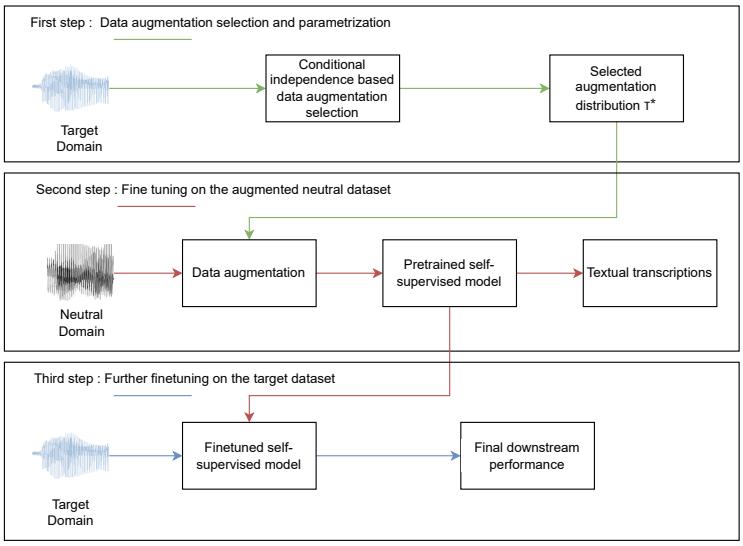

Figure 1: *Summary of the three steps of the method. 1. Starting from the target domain, an augmentation distribution is computed. 2. This distribution is used to distort a neutral dataset for a first fine-tuning. 3. A final fine-tuning is done on the target domain samples.*

recording conditions. From this distribution, we will sample augmentations to be applied to a larger "clean" dataset which will be used to fine-tune the SSL representations. The goal is to select augmentations bringing the "clean" dataset samples "closer" to those of the target domain, thus leading to better performance on its test sets. This section details the conditional-independence-based method developed to select a data augmentation distribution given the annotated target dataset. It starts by detailing the motivations behind the method, before delving into the technical details of the implementation.

2.1. Motivation and Technical Description

Motivation. Inspired by pretext-tasks selection for speech self-supervised learning, Zaiem *et al.* [22] have shown that conditional independence estimation may be used for automatic data augmentation in contrastive self-supervised learning settings. Furthermore, qualitative analysis has indicated that the distortions selected by this technique tend to be close to those of the target downstream dataset. Explicitly, applying a set of augmentations creates a set of augmented versions, often called "views", of the original samples. Minimizing, with respect to the augmentations selected, the dependence between the views and the IDs of the samples they originate from, conditionally on the downstream labels leads to a good choice of augmentations in contrastive learning settings.

Let us give an intuition about what happens in these conditional independence computations to understand why it can be useful for domain adaptation as well. Roughly, minimizing the conditional dependence described above maximizes, within the same downstream class, the invariance of distorted samples (*i.e.* views) to the ID of their original speech sample. If a given distortion (for instance, reverberation) is not present in any sample in the original target dataset, randomly applying this distortion would decrease in-class similarity. Inversely, applying augmentations already present in samples in the dataset makes it harder to distinguish their original samples' IDs given the distorted samples and, thus, lowers the conditional dependence estimator. Conditioning on the downstream labels retains the signal clues characterizing the downstream classes since it prevents selecting distortions that are only relevant to one class, as they would reduce in-class similarity in the other classes.

Technical Description. Precisely, let X and Y be respectively, a set of speech data points and their respective set of downstream labels which are in our case textual transcriptions. With τ an augmentation distribution from which one can sample a chain of augmentations, we compute a distorted dataset $X' = f(X, \tau)$, with f a function that randomly applies augmentations sampled from τ on the speech samples. Specifically, we can generate N augmented versions per speech sample to get the augmented set of data points X', with N a hyperparameter. Every sample x' in X' is a distorted version of a point x in the original dataset X. We will refer to the ID of the original point x as z, defining the Z set. The ID here corresponds to a discrete value indexing the speech segments X. In contrastive self-supervised learning settings [23], augmentation selection is crucial to incorporate the most relevant invariances in the learned representation into the downstream task of interest [24]. In this context, it has been shown that choosing the augmentation distribution τ that minimizes an estimator of the conditional dependence between X and Z given Y leads to the best downstream performance on speaker and language recognition tasks [22]. This work extends this approach in two manners, first applying it for domain adaptation in a supervised setting, and second extending it to the speech recognition task. We use for this the Hilbert-Schmidt Independence Criterion (HSIC) [25], a kernel-based dependence estimator, also validated on pretext task selection in previous works [4]. The lower the HSIC estimator, the more conditionally independent the two sets are and the better the augmentations should be.

In summary, to find the optimal augmentation distribution τ^*, we resort to minimizing the HSIC quantity with the augmented dataset $X' = f(X, \tau)$ according to $\tau^* = \arg\min_\tau HSIC(f(X, \tau), Z|Y)$

with $HSIC(X', Z|Y)$ an estimate of the conditional dependence between the distorted speech samples and their original IDs given their downstream textual labels.

2.2. Augmentation Distributions and Implementation

An augmentation distribution τ is characterized by a set of parameters that defines how the chain of augmentations is sampled

Table 1: *Augmentations, descriptions and parameter ranges*

Name	Description	Range (Unit)
Low Min	Lowpass minimal frequency cutoff	[100-500] (Hz)
Low Max	Lowpass maximal frequency cutoff	[1000-5000] (Hz)
High Min	Highpass minimal frequency cutoff	[1000,4000] (Hz)
High Max	Highpass maximal frequency cutoff	[4000,6000] (Hz)
Pitch min	Minimal pitch shift	[-6,-2] (semitones)
Pitch max	Maximal pitch shift	[2,6] (semitones)
Min SNR	Minimal SNR for coloured noise	[0,5] (dB)
Max SNR	Maximal SNR for coloured noise	[10,30] (dB)
Min Gain	Minimal gain	[-20,-10] (dB)
Max Gain	Maximal gain	[3,10] (dB)

during training and applied to the next data point. Precisely, every distribution $(\tau(p))_{1\leq p\leq P}$ is represented as a vector of $P = 17$ parameters representing either the probability of applying an augmentation or the boundaries of a uniform probability distribution used to sample the parameters of the augmentation (*e.g.* maximal signal-to-noise ratio value for noise addition).

Since the considered augmentations are not differentiable according to the considered parameters, we apply a random search to minimize the HSIC value described above. Thus, we sample random distributions and select the one with the lowest dependence scoring. Specifically, for every considered target dataset, we first sample $D = 100$ distribution parametrizations $(\tau_i)_{i\in[1,D]}$. For every parametrization τ_i, we compute the HSIC quantity following two steps. First, the augmented set $X'_i = f(X, \tau_i)$ is generated by computing $N = 20$ views of every speech sample in X. Then, $HSIC(X'_i, Z|Y)$ is computed following the technique described in [26]. For Y, we consider the 10 classes consisting of the 10 most used words in the dataset and take only the portion of the speech where the word is pronounced, using word-level forced alignment. The augmentation distribution with the lowest HSIC scoring is selected to be applied during fine-tuning.

3. Experiments

This section describes the experiments led to validate the proposed approach first in a simulated environment, then on real-world distorted datasets.

3.1. Shared Experimental Protocol

In all the experiments, the model is composed of two blocks: a pre-trained Wav2Vec2.0 Large model and a downstream decoder. The pre-trained model acts directly on the speech waveform and outputs an embedding of size $1,024$ every 20ms of speech. Two fully connected layers with a hidden size of $1,024$ map each frame vector to one of the considered characters. The whole model is fine-tuned using Connectionist Temporal Classification (CTC) [27] loss. During inference, greedy decoding is applied to the CTC probability outputs without any language-model-based re-scoring following the SpeechBrain recipe [21].

We employ the Torch-Audiomentations library from the Asteroid team [28] as it accelerates the computation of augmentations both during HSIC scoring and training. From the pool of available augmentations, we selected the ones that have demonstrated efficacy in enhancing recognition performance with the contrastive predictive coding method [8]. Hence, seven augmentations are considered: pitch shifting, reverberation, gain (which may reproduce clipping issues), colored noise addition, high and low pass filtering, and polarity inversion. The application of these distortions is controlled with a set of parameters listed in Table 1.

Table 2: *Mean WER results on distorted versions of LibriSpeech test splits. While scoring below the topline, our method, named "CI Augment", is significantly better than applying all or random augmentations. "Baseline" corresponds to an augmentation-free training.*

LS Split	Baseline	Random	CI Augment	Topline
test-clean	29.86	29.91	**27.20**	26.11
test-other	43.89	42.48	**40.68**	36.92

3.2. Oracle Experiment

Task-specific experimental protocol. In this part, a known distortion distribution is first applied to a clean testing set. The resulting data will be considered as the mismatching target domain (i.e. a simulated one). In a second time, using this generated "noisy" dataset, appropriate augmentations, selected using our conditional independence-based method, are applied to a clean training dataset that will be used for fine-tuning our self-supervised representations. As only the test set is distorted, this simulated experiment only involves one fine-tuning, contrarily to the real-data scenario, where a second fine-tuning stage is held on the target training data, as shown in Figure 1. This toy experiment has two advantages compared to a natural setting. First, it ensures that the distortions in the testing set can be replicated by the set of augmentations considered. Second, since we have access to the augmentation distribution that generated the "noisy" target dataset, it allows estimating the similarity between the augmentation distribution used to create the simulated testing domain and the one obtained with our method.

In these experiments, $A = 8$ augmentations distributions are sampled and applied on the LibriSpeech *test-clean* and *test-other* splits [29]. For every sampled distribution, these two distorted splits are then considered as the testing datasets. We apply the same augmentation distributions to the *dev-clean* and *dev-other* splits, and use these two sets to compute the optimal augmentations following the method described in the previous section. Finally, we use the computed distribution τ^* with the lower $HSIC$ estimator value as the augmentation for fine-tuning our SSL model on LibriSpeech *train-clean-100* split.

Results. Table 2 presents the results obtained on the test splits of LibriSpeech in the oracle experiments, with the column "CI Augment" (the name of the approach, CI standing for Conditional Independence) showing the results of the proposed approach. Each value corresponds to the mean of the values obtained with each of the A target datasets created with the sampled augmentation distributions. The "Topline" corresponds to the result obtained when the training samples are augmented using the same distribution as the one used to generate the distorted testing splits (i.e oracle scenario). Two baselines are considered: the first one referred to as "All" applies all the considered augmentations with the default parameters. Then, "Random" refers to the mean value obtained if applying the $(A-1)$ other toplines augmentation distributions. Our method, while performing worse than the topline, leads to a relative word error rate (WER) improvement of 12.7% compared to the baseline on *test-clean*.

This controlled experiment also enables us to verify if the selected augmentations result in acoustic conditions cloning, as suggested in Section 2.1. Indeed, the probabilities of applying

Table 3: *Mean WER results on distorted versions of LibriSpeech test-clean and test-other. Our method, named "CI Augment", outperforms the baselines and random augmentations for each one of the two contributors.*

Contributor	**Without Augmentations**			**With Augmentations**		
	train-clean-100	Contributor Only	*train-clean-100* + Contributor	All	Random	CI Augment
Contributor 1	102.52	73.0	27.71	27.95	27.33	**24.27**
Contributor 2	96.49	98.92	20.48	20.76	22.23	**16.49**

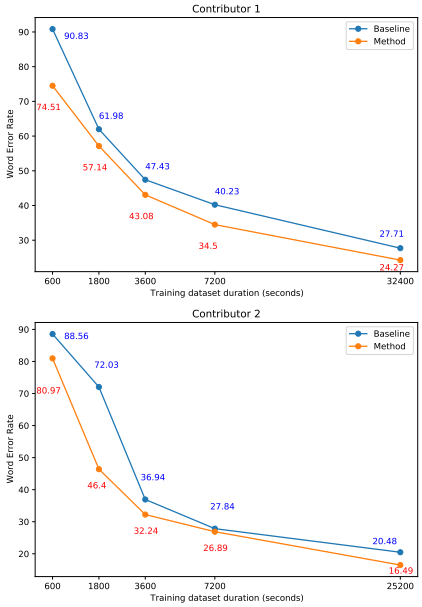

Figure 2: *Effect of selecting augmentations on the performance depending on the quantity of target domain training data for each of the two considered contributors. The x-axis is not linear.*

a given distortion to each testing set are known. To verify our intuition, for each one of the 8 augmentation distributions applied, we sample 200 other random augmentation distributions and score them using HSIC. For every scored distribution, we consider the vector composed by the seven probabilities of applying the considered distortions. Since these probabilities are known for the target distribution, we can compute an L_2 distance between the vector of probabilities of applying distortions used to create the target dataset, and those of the sampled scored distributions. We observe a Spearman correlation score of 0.51 between the HSIC scores and the distances between vectors of probabilities. Furthermore, the application probabilities of the 10 (top 5%) best scoring distributions are 15% closer to the target ones than those of the 10 worst scoring ones. These results indicate that the selected augmentations, *i.e.* those with low HSIC scoring, create samples closer to the target domain.

3.3. Experiments with Naturally Distorted Datasets

In this section, we test and validate the proposed approach on real low-resource "noisy" datasets.

Task-specific experimental protocol. The goal is to adapt a large clean "neutral" labeled dataset to better match the acoustic conditions of a small target dataset. The modified dataset is used during a first fine-tuning of the SSL representation, before further fine-tuning on the target dataset. To ensure a valid evaluation, the target dataset must meet two criteria: first, it should display consistent noisy recording and acoustic conditions. Second, neutral and target datasets should not exhibit different textual settings, *i.e* differences such as spontaneous versus read speech, as our augmentations only address acoustic distortions. The Librispeech *train-clean-100* is used as the clean dataset to be modified. The target datasets, on the other hand, correspond to the largest contributors of the CommonVoice 11.0 English dataset [30]. Starting from the ten most prolific contributors, two of them are finally selected after removing elements with heavy accents, and unintelligible or very clean recordings. For these two selected contributors, we partition the recorded samples into the train, validation, and test splits, and only use the training data to compute the augmentation distribution selection. The train splits are 9 and 7 hours long. More details can be found in the repository.

Results. Table 3 reports the WERs with or without augmentations during the first fine-tuning on *train-clean-100*. The first vertical part of the table shows the results obtained on the baselines without augmentations. "train-clean-100" corresponds to fine-tuning only on Librispeech *train-clean-100* split non-distorted. "Contributor Only" corresponds to training only on the contributor data. For all other columns, the model is fine-tuned on *train-clean-100* first, with or without augmentations, before further fine-tuning on the contributor data. The "CI Augment" column shows that the augmentations chosen with our conditional-independence-based method lead to better target performance than applying no, all, or random augmentations on the neutral training split. The relative improvement compared to the augmentation-free baseline reaches 12.4% for Contributor 1 and 19.5% for Contributor 2.

Furthermore, we study how this affects the amount of target domain data needed (see Figure 2). We start by fine-tuning with the chosen distortions for the "Method" lines and on the clean original LibriSpeech dataset for the "Baseline" lines. Then, the duration of annotated target data used is augmented gradually . For the two contributors, the orange curve representing the evolution of the WER after fine-tuning with the computed distortions is always below the blue curve corresponding to the baseline. The effect is particularly visible with Contributor 1 with a performance 16.6% higher relatively when training with only 2 hours.

4. Acknowledgements

This work has benefited from funding from l'Agence de l'Innovation de Défense, and was performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011012801R1).

5. Conclusion

Self-supervised representations severely underperform when facing acoustic domain mismatch. We have introduced a method using automatic data augmentation selection to reduce the drop in performance when switching of acoustic domains. Experiments led in controlled and natural settings validate our assumption and method, and also show that it helps reduce the quantity of annotated data needed in the target domain.

6. References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[2] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[3] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," 2020.

[4] S. Zaiem, T. Parcollet, and S. Essid, "Pretext tasks selection for multitask self-supervised speech representation learning," 2021.

[5] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating the reliability of acoustic speech embeddings," in *INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association*, Shanghai / Vitrtual, China, Oct. 2020.

[6] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," 2020.

[7] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *Interspeech 2020*. ISCA, oct 2020.

[8] 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)*, 2021, pp. 215–222.

[9] A. Sriram, M. Auli, and A. Baevski, "Wav2Vec-Aug: Improved self-supervised training with limited data," pp. 4950–4954, jun 2022.

[10] M. Riviere, J. Copet, and G. Synnaeve, "ASR4REAL: An extended benchmark for speech models," oct 2021.

[11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, oct 2021.

[12] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," pp. 721–725, apr 2021.

[13] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities." Dublin, Ireland: Association for Computational Linguistics, 2022.

[14] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Esteve, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech," 2021.

[15] M. Olvera, E. Vincent, and G. Gasso, "On The Impact of Normalization Strategies in Unsupervised Adversarial Domain Adaptation for Acoustic Scene Classification," in *ICASSP*, vol. 2022-May, may 2022, pp. 631–635.

[16] T. Tanaka, R. Masumura, H. Sato, M. Ihori, K. Matsuura, T. Ashihara, and T. Moriya, "Domain Adversarial Self-Supervised Speech Representation Learning for Improving Unknown Domain Downstream Tasks," 2022.

[17] V. S. Lodagala, S. Ghosh, and S. Umesh, "Pada: Pruning assisted domain adaptation for self-supervised speech representations," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 136–143.

[18] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, "Self Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, apr 2022.

[19] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5834–5838.

[20] K. P. Huang, Y.-K. Fu, Y. Zhang, and H.-y. Lee, "Improving distortion robustness of self-supervised speech processing tasks with domain adaptation," 2022.

[21] 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.

[22] S. Zaiem, T. Parcollet, and S. Essid, "Automatic Data Augmentation Selection and Parametrization in Contrastive Self-Supervised Speech Representation Learning," apr 2022.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[24] T. Xiao, X. Wang, A. Efros, and T. Darrell, "What Should Not Be Contrastive in Contrastive Learning," 2020.

[25] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola, "A kernel statistical test of independence," 01 2007.

[26] S. Zaiem, T. Parcollet, and S. Essid, "Conditional independence for pretext task selection in self-supervised speech representation learning," in *Interspeech 2021*. ISCA, aug 2021.

[27] 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*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376.

[28] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 2637–2641, may 2020.

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

[30] 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," 2020.