Automatic data augmentation for training and adaptation of speech self-supervised models

Salah Zaiem, Titouan Parcollet and Slim Essid
salah.zaiem@telecom-paris.fr

INTERSPEECH 2022/2023
Atelier DSAIDIS - 24/05/2023

ADASP audio data analysis signal processing
TELECOM Paris
IP PARIS
Samsung AI Center-Cambridge

Outline

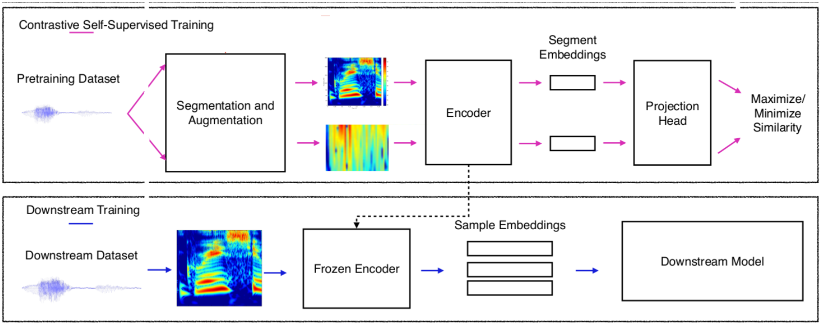
Introduction

Augmentations Selection and Parametrisation

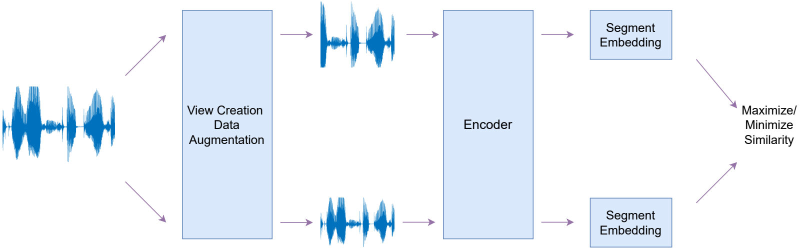
Testing Procedure and Results

Automatic Data Augmentation for Domain Adaptation

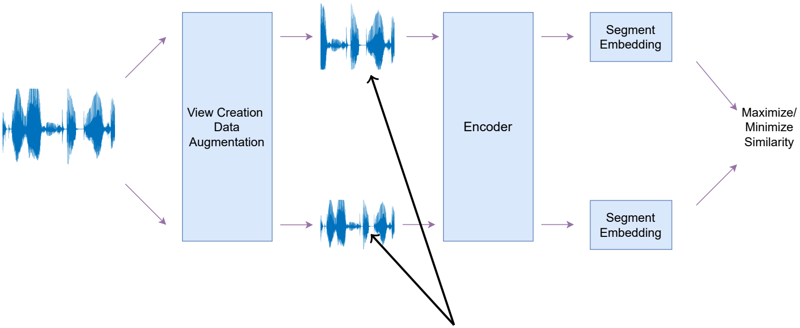
Self-Supervised Learning



Contrastive Learning

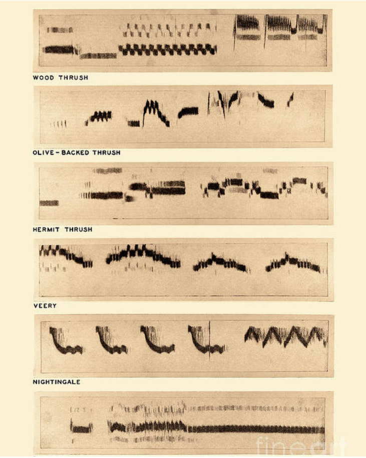


Contrastive Learning

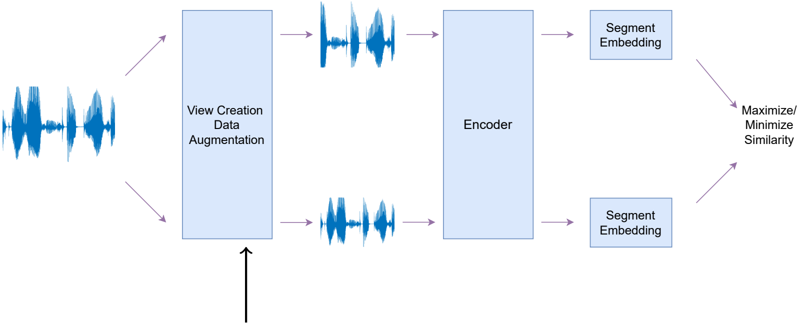


Should share the same downstream label !

Contrastive Learning



Contrastive Learning



Given the downstream task,
how to select and parametrize
the data augmentations ?

Related Works

- Saeed, A., Grangier, D., Zeghidour, N. / Contrastive Learning of General-Purpose Audio Representations. (IEEE Signal Processing Letters 2021) $\longrightarrow$ Without augmentations baseline.
- Xiao, T., Wang, X., Efros, A. A., Darrell, T. / What Should Not Be Contrastive in Contrastive Learning. (ICLR 2021). $\longrightarrow$ First diagnostic of the issue in computer vision.
- Chavhan, R., Gouk, H., Stuehmer, J., Heggan, C., Yaghoobi, M., Hospedales, T. / Amortised Invariances for Contrastive Self-Supervision. (ICLR 2023). $\longrightarrow$ Selects the relevant invariances during the fine-tuning phase.

Outline

Conditional Independence based estimator

Self supervised learning : learning representations through solving pretext tasks.

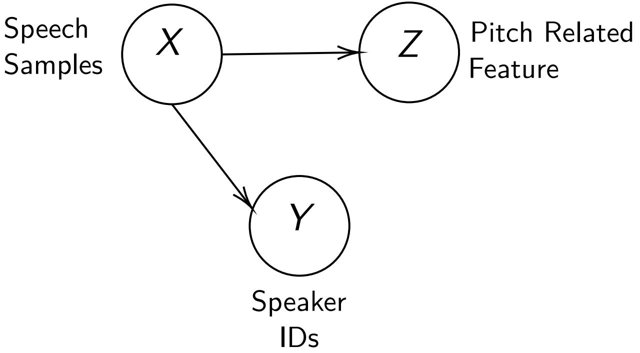
Previous work

Speech samples $\perp$ Pretext task labels $|$ Downstream labels
$\longrightarrow$ Good pretext task

S. Zaiem *and al.*, "Pretext Tasks Selection for Multitask Self-Supervised Audio Representation Learning," in IEEE JSTSP, 2022

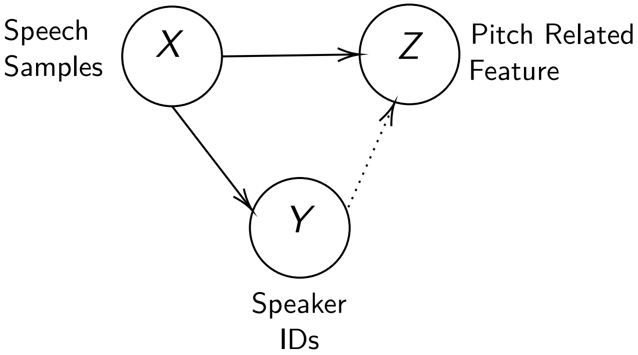
Conditional Independence based estimator

Speech samples $\perp$ Pretext task labels | Downstream labels



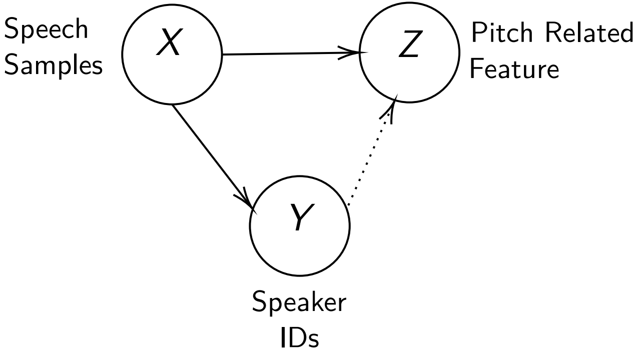
Conditional Independence based estimator

Speech samples $\perp$ Pretext task labels | Downstream labels



Conditional Independence based estimator

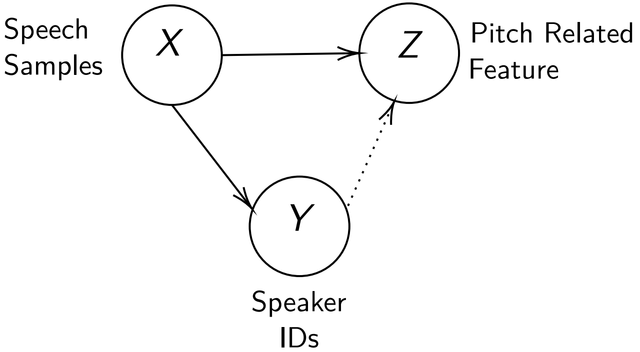
Speech samples $\perp$ Pretext task labels | Downstream labels



Conditional dependence estimate : $HSIC(X, Z|Y)$

Conditional Independence based estimator

Speech samples $\perp$ Pretext task labels | Downstream labels



Conditional dependence estimate : $HSIC(X, Z|Y)$

How is pretext task selection related to contrastive learning?

Link with Contrastive Learning

Key observation

Contrastive learning $\approx$ Task of retrieving the original speech sample from an augmented version (view)

If we can retrieve the ID of the original sample, we can maximise the similarity between two generated segments.

Link with Contrastive Learning

Key observation

Contrastive learning $\approx$ Task of retrieving the original speech sample from an augmented version (view)

- Inputs : pretraining dataset X_{unl}, augmentation distribution τ
- Creating the views : $X_{unl} \xrightarrow{f_\tau(x)} X'_{unl}$
- Contrastive learning can be seen as as the task Z_τ consisting for an augmented point x' in retrieving the ID of $f_\tau^{-1}(x')$

Link with Contrastive Learning

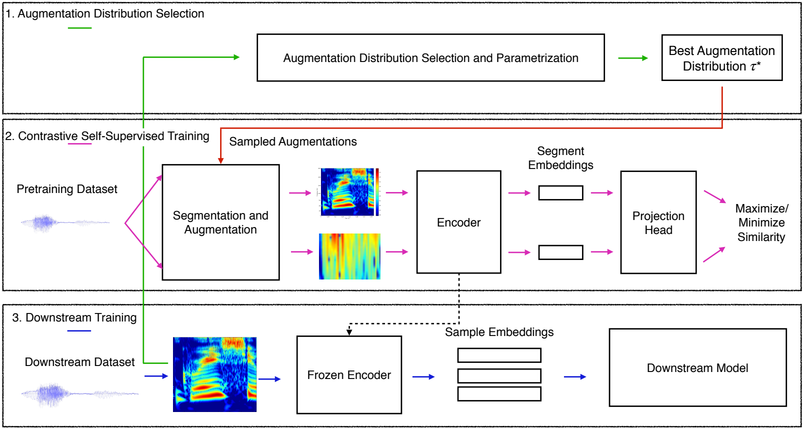
- Contrastive learning pretraining is now seen as solving task Z_τ.
- The lower the $HSIC(X, Z|Y)$ (the conditional indpendence estimator), the better is the the pretraining task.

Link with Contrastive Learning

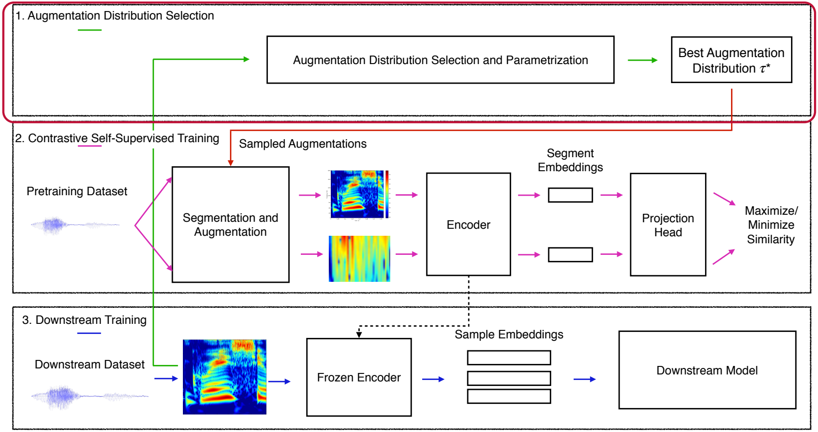
- Contrastive learning pretraining is now seen as solving task Z_τ
- The lower the $HSIC(X, Z|Y)$ (the conditional indpendence estimator), the better is the the pretraining task
- For a given task (X, Y), τ is chosen such as :

$$\tau^* = \arg\min_{\tau} HSIC(X, Z_\tau | Y)$$

Three steps validation



First step



1. Augmentation Distribution Selection
Augmentation Distribution Selection and Parametrization
Best Augmentation Distribution τ^*
2. Contrastive Self-Supervised Training
Sampled Augmentations
Pretraining Dataset
Segmentation and Augmentation
Encoder
Segment Embeddings
Projection Head
Maximize/ Minimize Similarity
3. Downstream Training
Downstream Dataset
Frozen Encoder
Sample Embeddings
Downstream Model

Selecting the Augmentation Distribution

An augmentation distribution τ is defined by a set of parameters defining how a chain of augmentations is sampled during pretraining

Set of considered augmentations :

- Reverberation
- Band Scaling
- Pitch Shifting
- Clipping
- Timedropping

Selecting the Augmentation Distribution

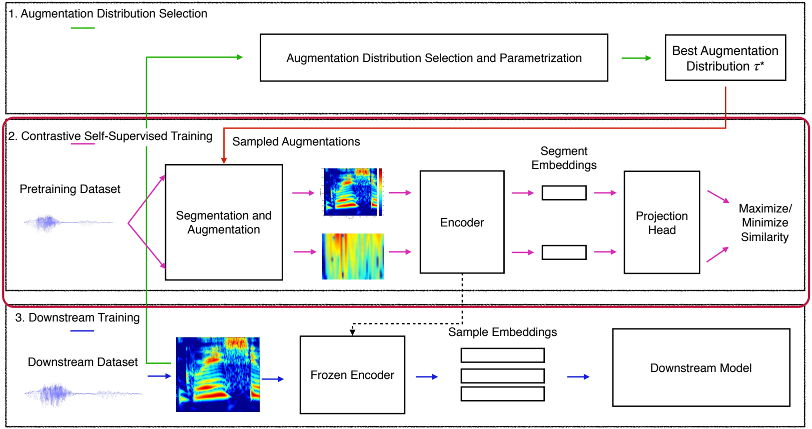
Every distribution τ is represented as a vector of $P = 14$ parameters
Probabilities of applying an augmentation / controlling parameters

Name	**Description**	**Range**
Room scale min	Min room size	[0,30]
Room scale max	Max room size	[30,100]
Band Scaler	Scales the rejected band	[0,1]
Pitch Shift Max	Amplitude of a pitch shift	[150,450]
Pitch Quick pr.	Speeds pitch shifting	[0,1]
Clip Min	Minimal clip factor	[0.3, 0.6]
Clip Max	Maximal clip factor	[0.6, 1]
Timedrop max	Size of a time dropout	[30-150] ms

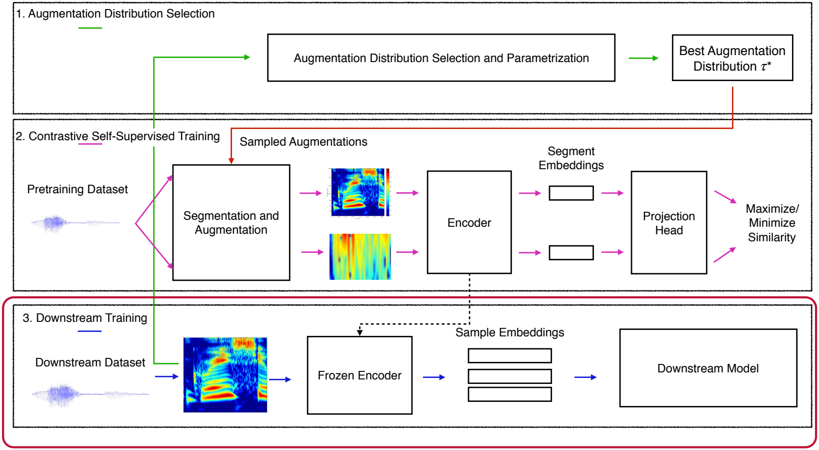
To minimize the described HSIC, we resort to a random search among the parameters

Outline

Next steps : Pretraining



Next steps : Finetuning



Datasets

Datasets Roles and Descriptions

Task	Dataset	$\sim$Dur.(train)	Speak./Lang.
Pretraining	CommonVoiceEn6.1	1686 hours	$\sim$66173
Lang. ID	VoxForge	176 438 utt	6
Speak Reco	VoxCeleb1	148 642 utt	1251

Architecture details very close to COLA, our baseline, for pretraining. And finetuning according to the SUPERB benchmark of SSL representations.

Downstream Results

All (Default) : applying on every point all the augmentations with default parameters.
Random : mean of 5 runs with randomly sampled distributions.

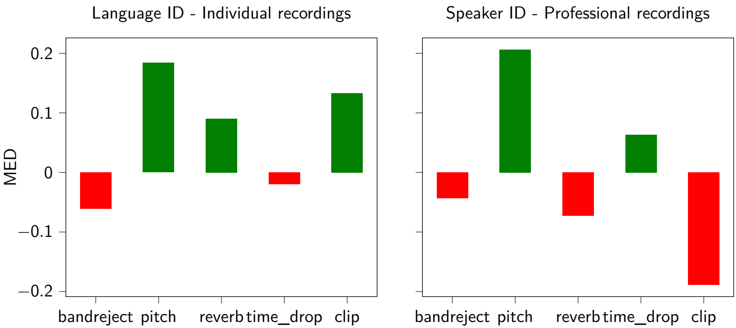
Down. Task	**COLA**	**Our Implementations**			
		Without	Random (5 runs)	All (Default)	Selected
Language ID	71.3	76.1	84.9	84.3	**87.1**
Speaker ID	29.9	35.2	32.0	45.1	**47.8**

Qualitative analysis

Considered quantity (MED): Difference of the probability of picking an augmentation between the best and worst scoring augmentations, depending on the downstream dataset.

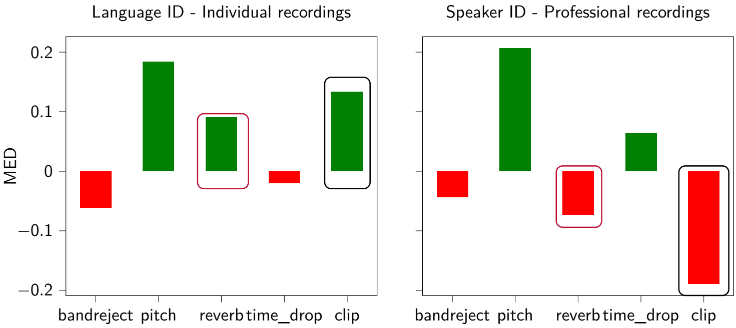
Qualitative analysis

Considered quantity (MED): Difference of the probability of picking an augmentation between the best and worst scoring augmentations, depending on the downstream dataset.



Qualitative analysis

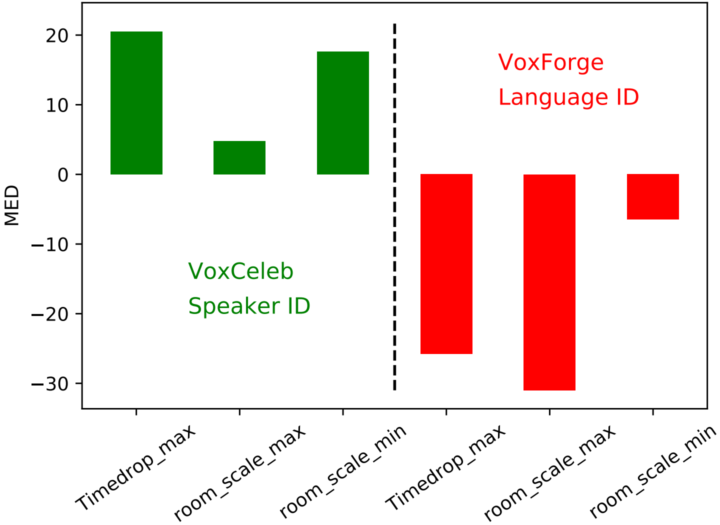
Considered quantity (MED): Difference of the probability of picking an augmentation between the best and worst scoring augmentations, depending on the downstream dataset.



Recording conditions seem to prevail in selecting the relevant augmentations.

Qualitative analysis

Differences in parameters values :



First Conclusions

Given a downstream task, can we choose the augmentations for a contrastive learning based pretraining ?

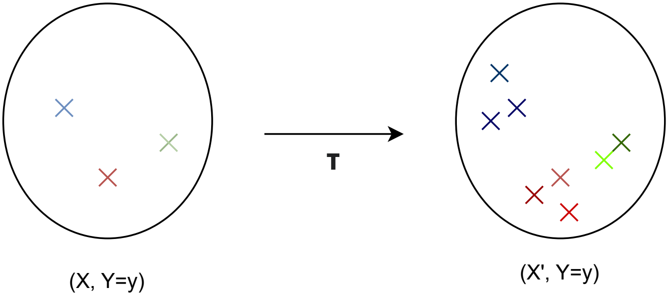
First Conclusions

Given a downstream task, can we choose the augmentations for a contrastive learning based pretraining ?

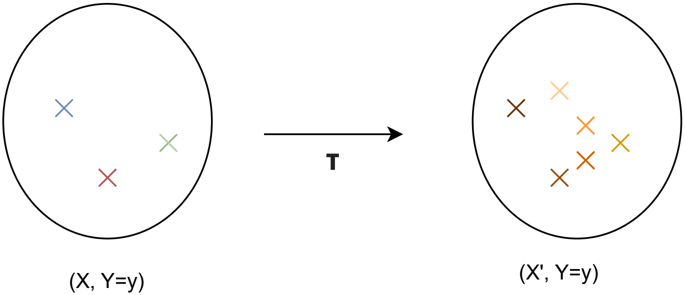
- Conditional independence based data augmentation selection and parametrization

Outline

Some more intuition



Some more intuition



- Collapse is prevented by fixing limits to the augmentations sampled.
- Conditinoning on the downstream classes allows keeping discriminative signal clues.

Acoustic conditions cloning

Selected augmentations seem to replicate non downstream class dependent distortions.
$\longrightarrow$ Could be used to clone acoustic distortions within a target dataset.

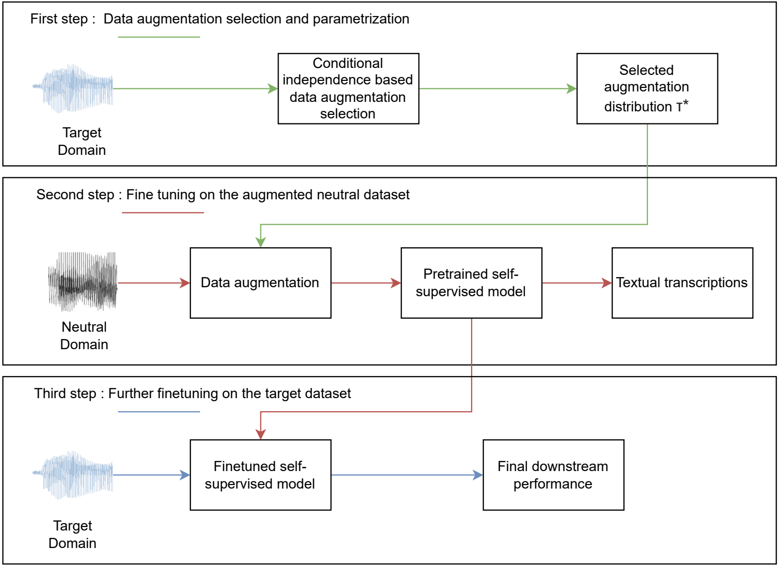
Domain adaptation for Self-Supervised Models

When encountering audios with conditions very different from pretraining ones, SSL models suffer high ASR performance drops.
Fine-tuning on 10 hours from LibriSpeech $\rightarrow$ 8% WER.
Fine-tuning on 10 hours from CommonVoice $\rightarrow > 20\%$ WER.

Setting and Idea

- Annotated target dataset D_T, small size, and specific acoustic conditions.
- Large clean ("Neutral") dataset D_C
- Compute the augmentation policy τ_T using the conditional-independence based approach. Apply it on D_C to obtain D_{CT}. Then fine-tune on D_{CT} as a first fine-tuning step.

Setting and Idea



Settings

- SSL Model : Wav2vec 2.0 Large
- Downstream head : linear decoder trained with CTC Loss. No language modelling for decoding
- Downstream classes : most common words. We cut at word level using forced alignment.

Oracle Experiments

We apply a known augmentation distribution to the test splits of Librispeech and use it as target domain. (10 times)

- Sample an augmentation distribution τ
- Apply τ on LibriSpeech *test-clean* to create the testing set and on *dev-clean* to create D_T.
- Apply our method on D_T to obtain a distribution τ^*
- Apply τ^* on LibriSpeech *train-clean-100* to obtain D_{CT}, then use D_{CT} for training.

Oracle Experiments

Table: Mean WER results on distorted versions of LibriSpeech test-clean and test-other.

Split	No Aug	Baseline	Random	CI Augment	Topline
test-clean	33.81	29.86	29.91	**27.20**	26.11
test-other	44.12	43.89	42.48	**40.68**	36.92

- Baseline : All augmentations with default parameters.
- Random : Mean of 9 runs with random augmentation policies during training.
- **Topline** : Applying the test distortions on the train data.

Experiments on natural datasets

Conditions for the datasets :

- Small target distorted dataset => interesting challenge
- Textual correspondance with the neutral dataset (read speech)
- Coherent and regular acoustic conditions.

We took the samples from the most productive contributors (contributors) of CommonVoice 11.0, and selected two contributors respecting the conditions.

Experiments on natural datasets

Two steps fine-tuning, first on distorted LibriSpeech *train-clean-100*, and second on the contributor data.

Contributor	Without Augmentations			With Augmentations		
	Train100	Only Contrib.	Train100 + Contrib.	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**

Table: WER Results on the two considered CommonVoice contributors.

Varying the available annotation

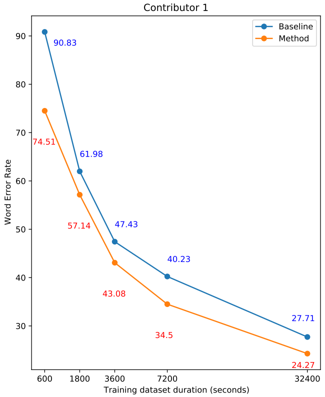


Figure: Effect of choosing suitable augmentations on the performance depending on the quantity of in-domain training data for each of the two considered contributors.

Conclusion

- Our approach allows for automatic data augmentation for "faster" adaptation of self-supervised speech models.
- Needs to see the effect of bigger "Neutral" datasets.
- Limited to acoustic mismatch.

Thank You

Thank you all for your attention !

Please feel free to ask any question