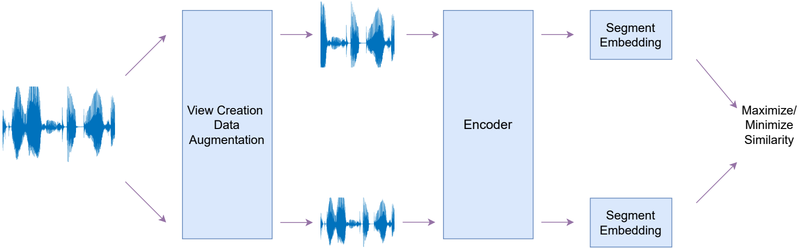
Automatic Data Augmentation Selection and Parametrization in Contrastive Self-Supervised Speech Representation Learning

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

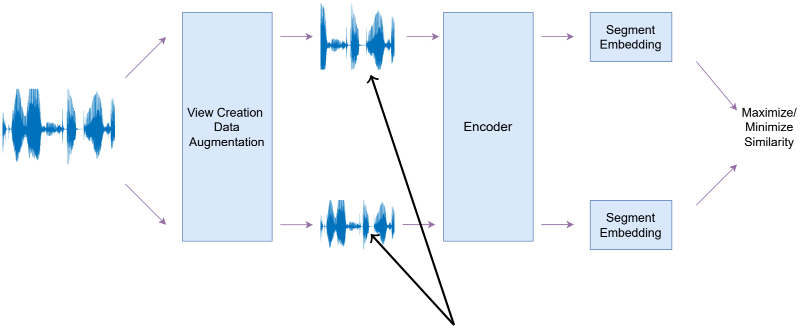
INTERSPEECH 2022

ADASP audio data analysis signal processing
TELECOM Paris
IP PARIS
LABORATOIRE INFORMATIQUE D'AVIGNON

Contrastive Learning

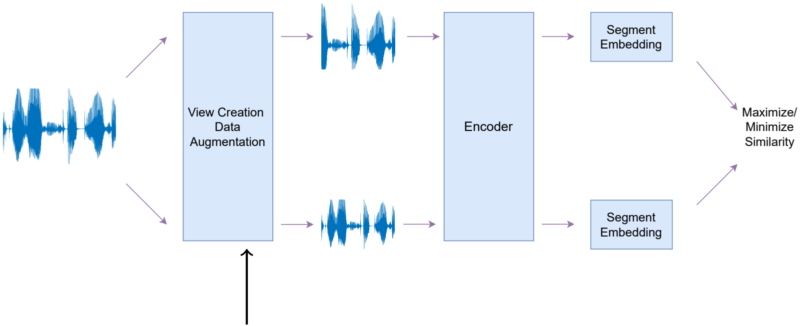


Contrastive Learning



Should still share the same downstream label !

Contrastive Learning



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

Related works

- A. Saeed, D. Grangier, and N. Zeghidour, Contrastive learning of general-purpose audio representations, 2020.
- H. Al-Tahan and Y. Mohsenzadeh, CLAR: Contrastive Learning of Auditory Representations, AISTATS, 2021.
- T. Xiao, X. Wang, A. Efros, and T. Darrell, What Should Not BeContrastive in Contrastive Learning, 2021.

Outline

Outline

Conditional Independence based estimator

Self supervised learning : learning representations through solving pretext tasks.

Previous work

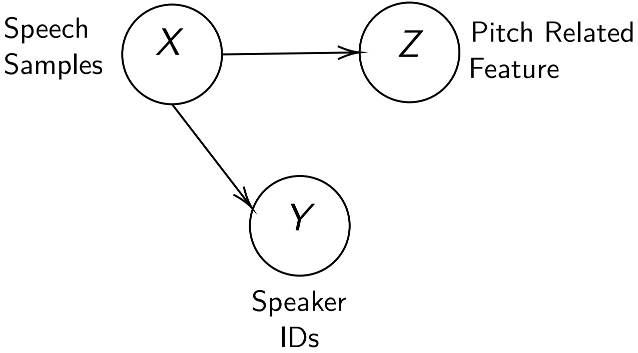
Speech samples $\perp$ Pretext task labels $|$ Downstream labels

$\longrightarrow$ Good pretext task

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

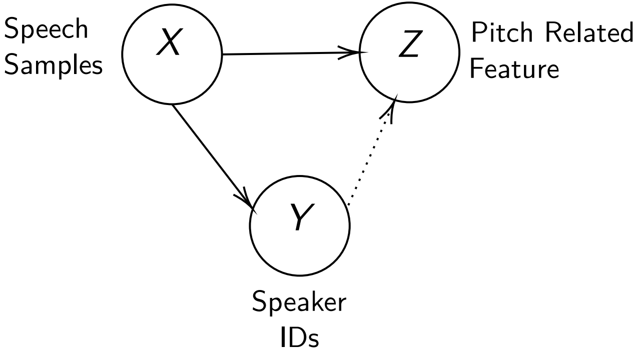
Conditional Independence based estimator

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



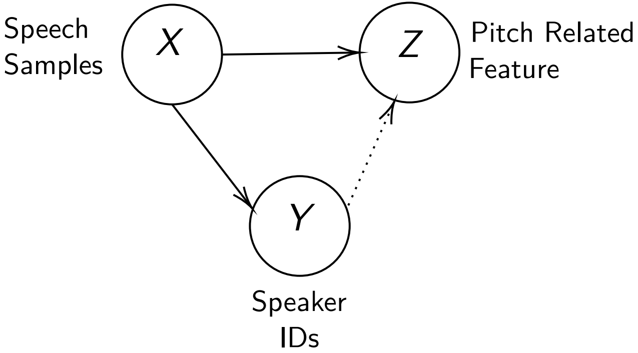
Conditional Independence based estimator

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



Conditional Independence based estimator

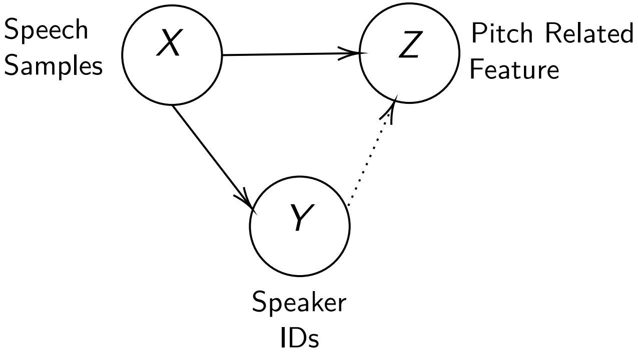
Speech samples $\perp$ Pretext task labels | Downstream labels



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 this 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 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

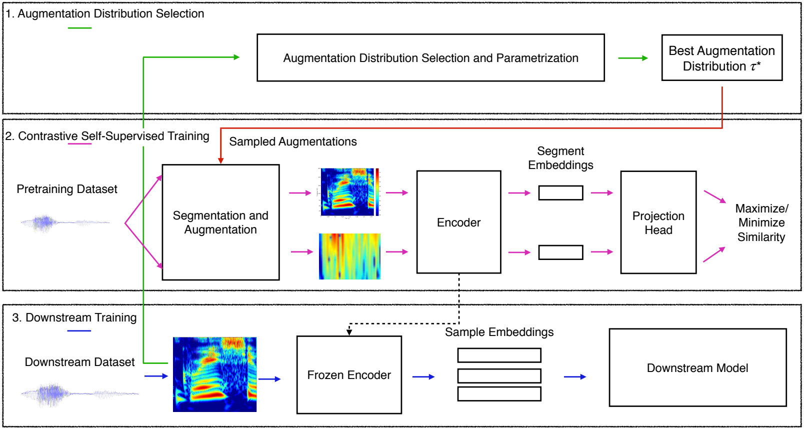
- Contrastive learning pretraining is now seen as solving task Z_τ.
- The lower the $HSIC(X, Z|Y)$ (the conditional indpendence estimator), the better is the the pretraining task.

Link with Contrastive Learning

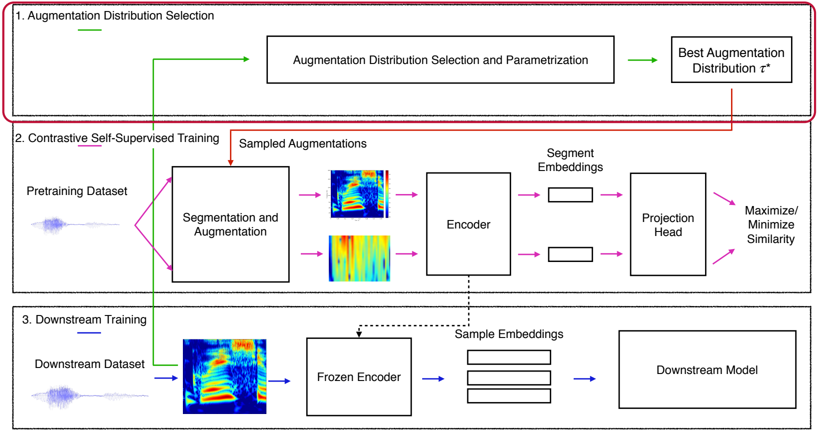
- Contrastive learning pretraining is now seen as solving task Z_τ
- The lower the $HSIC(X, Z|Y)$ (the conditional indpendence estimator), the better is the the pretraining task
- For a given task (X, Y), τ is chosen such as :

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

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

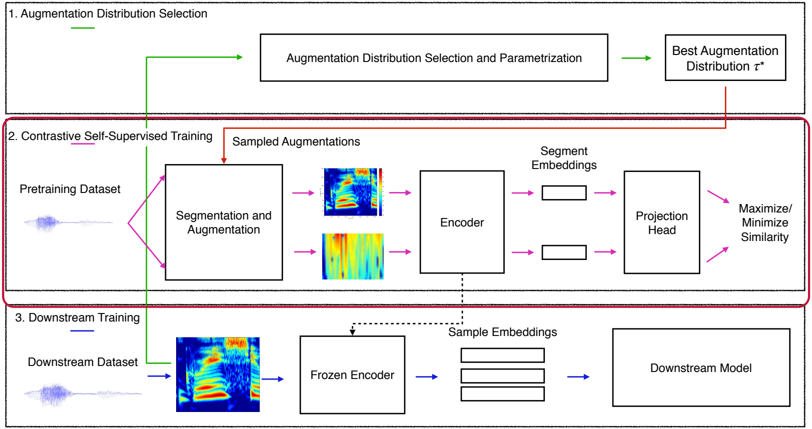
Outline

Introduction

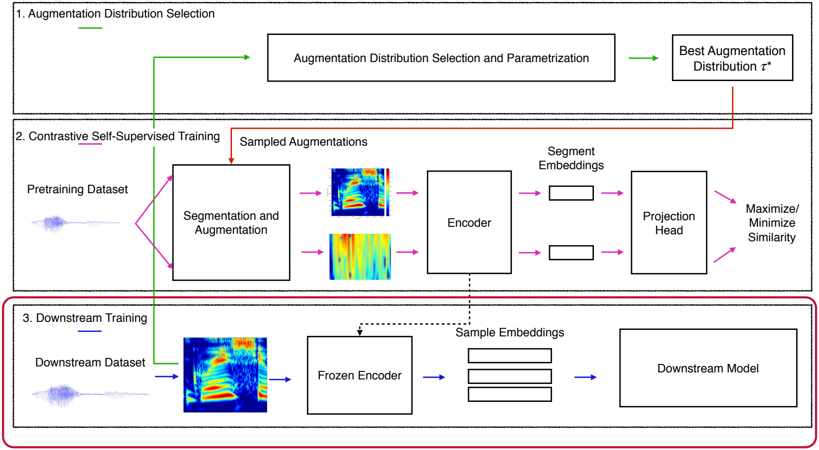
Augmentations Selection and Parametrisation

Testing Procedure and Results

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.

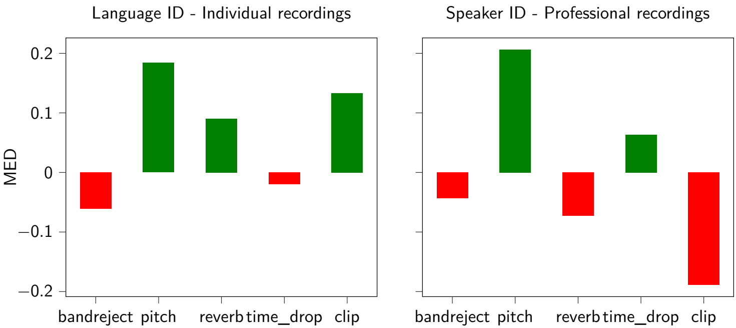
Down. Task	**COLA**	**Our Implementations**			
		Without	Random (5 runs)	All (Default)	Selected
Language ID	71.3	76.1	84.9	84.3	**85.2**
Speaker ID	29.9	35.2	32.0	45.1	**46.9**

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.

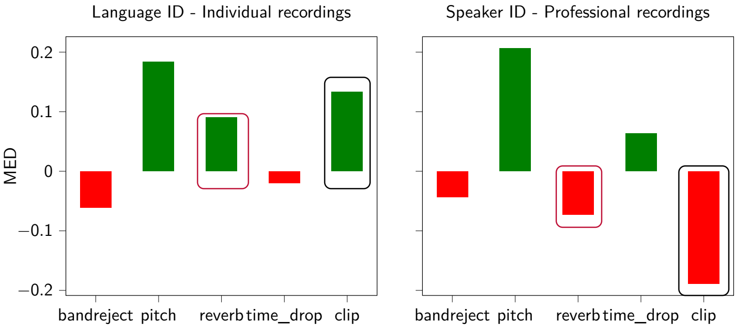
Qualitative analysis

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

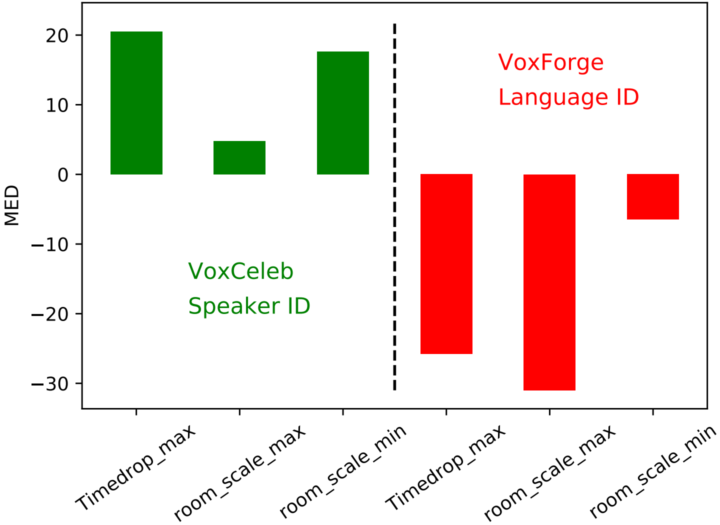
Considered quantity (MED): Difference of the probability of picking an augmentation between the best and worst scoring augmentations, depending on the downstream dataset.



Recording conditions seem to prevail in selecting the relevant augmentations.

Qualitative analysis

Differences in parameters values :



Conclusion

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

Conclusion

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

- Conditional independence based data augmentation selection and parametrization
- Further works on data augmentation in supervised settings

Thank You

Thank you all for your attention !

Please feel free to ask any question