Pretext Tasks Selection for Multitask Self-Supervised Speech and Audio Representation Learning

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

AAAI 2022
The 2nd Workshop on Self-supervised Learning for Audio and Speech Processing

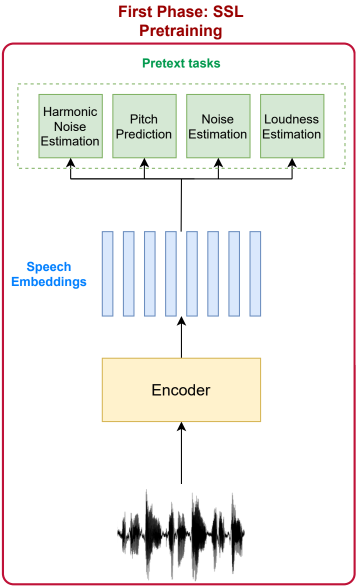
Outline

Introduction

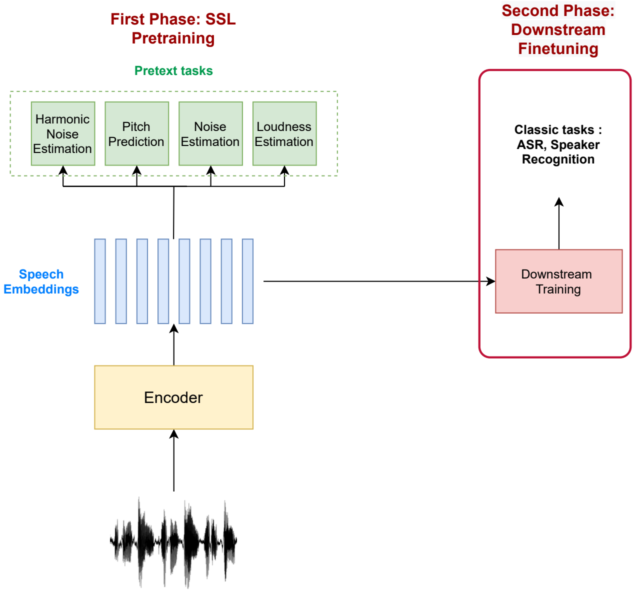
Conditional Independence (CI) Based Estimator

Multitask Self-supervised Learning

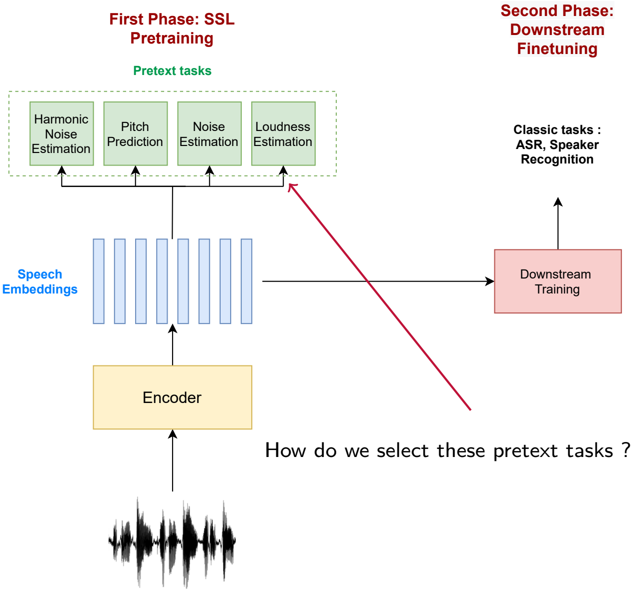
Introduction



Introduction



Introduction



How do we select these pretext tasks ?

Objective

How do we select the self-supervised pretext tasks optimally towards solving a given downstream one ?
Can we find a function scoring the usefulness of a given pretext task towards solving a downstream one ?

Outline

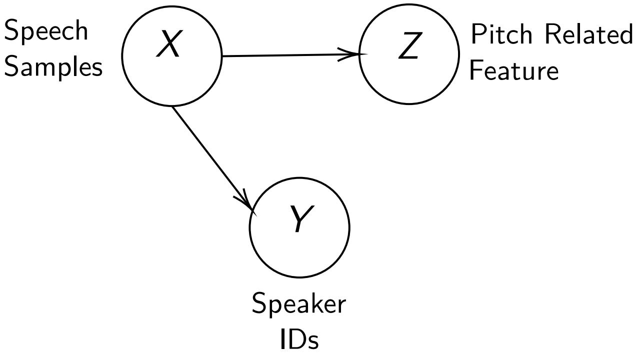
Conditional Independence based estimator

Main Idea

Speech samples $\perp$ Pretext task labels (Pseudo-labels) | Downstream labels
$\longrightarrow$ Good pretext task.

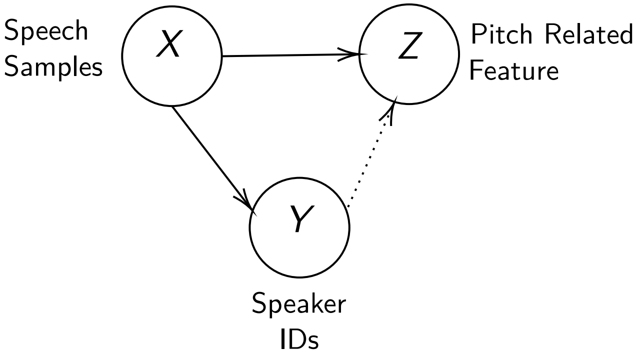
Conditional Independence based estimator

Speech samples $\perp$ Pseudo labels | Downstream labels



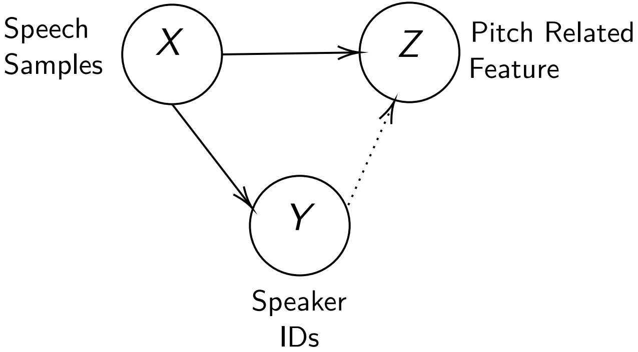
Conditional Independence based estimator

Speech samples $\perp$ Pseudo labels | Downstream labels



Conditional Independence based estimator

Speech samples $\perp$ Pseudo labels | Downstream labels



Non trivial to compute.

Hilbert Schmidt Independence Criterion (HSIC)

- Zaiem, S., Parcollet, T., Essid, S. (2021). Conditional independence for pretext task selection in Self-supervised speech representation learning. INTERSPEECH 2021.
- Kernel-based independence testing between speech samples and pseudo labels

Hilbert Schmidt Independence Criterion (HSIC)

- Zaiem, S., Parcollet, T., Essid, S. (2021). Conditional independence for pretext task selection in Self-supervised speech representation learning. INTERSPEECH 2021.
- Kernel-based independence testing between speech samples and pseudo labels

$$HSIC(X,Z|Y) = \frac{1}{M}\sum_{c\in\mathscr{C}} HSIC_c(X,Z) \times n_c.$$

Outline

Multi-task SSL

Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., Bengio, Y. (2019). Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks.
Doersch, C., Zisserman, A. (2017). Multi-task Self-Supervised Visual Learning.

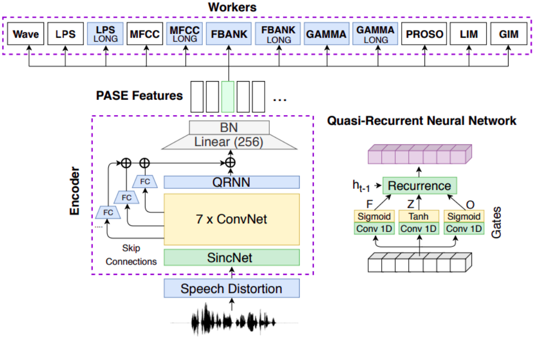


Fig. 1. The proposed PASE+ architecture for self-supervised learning. In blue are the main differences with the previous version of PASE.

Multi Pretext Tasks Selection

From individual pretext task selection to multi-tasked self supervised representation learning

Multi Pretext Tasks Selection

From individual pretext task selection to multi-tasked self supervised representation learning

And if we learn a group simultaneously, how do we weight the corresponding losses ?

New Problem

Best regrouping pretext task $Z_\lambda = (\lambda_1 Z_1, ..., \lambda_k Z_k)$ with :

- $(Z_i)_{i \in [0,k]}$ the individual pretext tasks
- $(\lambda_i)_{i \in [0,k]}$ the weights corresponding to their losses during the pretraining phase.

New Problem

Best regrouping pretext task $Z_\lambda = (\lambda_1 Z_1, ..., \lambda_k Z_k)$ with :

- $(Z_i)_{i \in [0,k]}$ the individual pretext tasks
- $(\lambda_i)_{i \in [0,k]}$ the weights corresponding to their losses during the pretraining phase.

Constraints on the weights :

- Positive weights (non adversarial learning)
- Not too low $=>$ constant sum.
- Sparse weighting vector.

Constraints on the weights

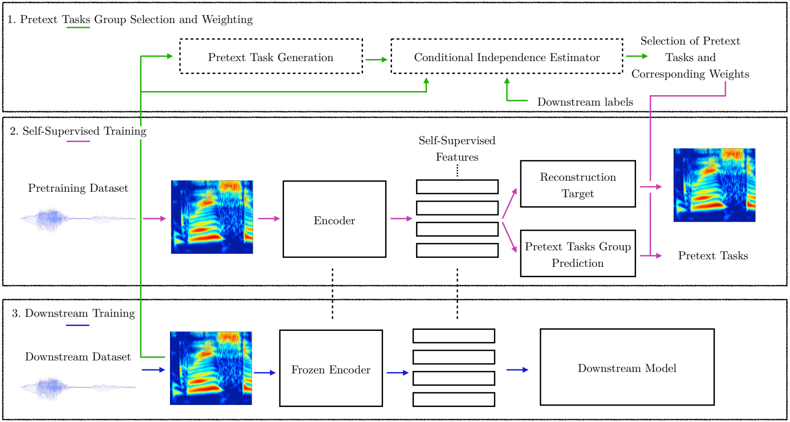
- Positive weights (non adversarial learning)
- Not too low => constant sum.
- Sparse weighting vector.

$$\min_{W \in \mathbb{R}^k} \quad HSIC(Z_\lambda, X|Y), \text{ s.t. } \lambda = f(W),\ Z_\lambda = (\lambda_1 Z_1, ..., \lambda_k Z_k). \tag{1}$$

with f in [Softmax, Sparsemax].

Martins, A. F. T., Astudillo, R. F. (2016). From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification.

Validation steps



Pretext tasks: pseudo-labels prediction

Candidate pseudo-labels and descriptions

Pseudo-label	Description
Loudness	Intensity & approx. loudness
F0	Fundamental Frequency
Voicing	Voicing Decision
Alpha Ratio	Ratio of spectrum intensity % 1000 Hz
Zero Crossing Rate	Zero crossing number per frame
RastaSpec L1Norm	L1 Norm of Rasta Spectrum
log HNR	log of Harmonicity to Noise Ratio

Datasets

Datasets Roles and Descriptions

Task	Dataset	$\sim$Dur.(train)	Speakers
Speech			
Pretraining	CommonVoiceEn6.1	1686 hours	$\sim$66173
ASR	Libri100	100 hours	251
Speak Recog.	VoxCeleb1	148642 utt	1251
Emotion Recog.	IEMOCAP	12 hours	10
Music			
Music Pretrain.	Audioset(Music Inst.)	155 hours	Irr.
Solo Instr.	Medley-solos-DB	18 hours	Irr.
Multi Instr.	OpenMIC-2018	55 hours	Irr.

First Results

$$L_{SSL} = MSE_{mel} + MSE_{mfcc} + \sum_{i=1}^{k} \lambda_i \ell_1(Z_i), \qquad (2)$$

Table: Results observed with the proposed selection strategies on the three considered downstream tasks.

Models	**LibriSpeech** *(WER % ↓)*		**VoxCeleb1** *(EER % ↓)*	**IEMOCAP** *(Acc % ↑)*
	No LM	*LM*		
PASE+ (Ravanelli, 2020)	25.11	16.62	11.61	57.86
Selections				
All	21.98 ± 0.36	11.70 ± 0.27	11.90± 0.32	56.4 ± 1.3
MRMR	18.94 ± 0.34	10.36 ± 0.26	10.56 ± 0.31	59.6 ±1.29
RFE	20.02 ± 0.34	11.42 ± 0.27	11.91 ± 0.33	55.8 ± 1.3
Softmax	**13.17± 0.28**	**8.00** ± **0.23**	9.24 ± 0.29	60.6 ± 1.27
Sparsemax	17.18 ± 0.32	10.41 ± 0.26	**8.63** ± **0.27**	**60.8** ± **1.28**

Extending wav2vec 2.0

Effect of adding carefully selected pretext tasks to a powerful CPC task ?

$$L_{SSL} = L_{W2V} + \sum_{i=1}^{k} \lambda_i \ell_1(Z_i). \tag{3}$$

Extending wav2vec 2.0

Effect of adding carefully selected pretext tasks to a powerful CPC task ?

$$L_{SSL} = L_{W2V} + \sum_{i=1}^{k} \lambda_i \ell_1(Z_i). \tag{4}$$

Table: Results observed retraining the Wav2vec2 model with and without weighted pretext tasks using the sparsemax method. “Fr." and “Fine." also respectively refer to Frozen and Finetuned settings.

Selections	**LibriSpeech** *(WER % ↓)*		**VoxCeleb1** *(EER % ↓)*		**IEMOCAP** *(Acc % ↑)*	
	Fr.	Fine.	Fr.	Fine.	Fr.	Fine.
wav2vec 2.0 *BASE*	17.93 ± 0.33	10.21 ± 0.25	7.20 ± 0.26	5.35 ± 0.22	56.6 ± 1.2	**74.0 ± 1.16**
wav2vec 2.0 *BASE* + Naive selection	17.23 ± 0.32	10.10 ± 0.25	6.80 ± 0.25	**5.05 ± 0.21**	57.4 ± 1.3	73.7 ± 1.16
wav2vec 2.0 *BASE* -Sparsemax	**16.70 ± 0.31**	**9.18 ± 0.24**	**6.57 ± 0.25**	5.30 ± 0.22	**59.5 ± 1.29**	**74.0 ± 1.16**

Task change : Musical Instrument Recognition

Table: Results observed with the proposed selection strategies on the two considered downstream instrument recognition tasks. Accuracy on the test set is computed for Medley-solos-DB while mean F1 Score is shown for OpenMIC. Higher is better for both.

Models	**Medley-solos** *(Acc% ↑)*	**OpenMIC-2018** *(mean-F1 ↑)*
PASE+ (Ravanelli, 2020)	None	64.1
Selections		
All	66.2 $\pm$ 0.83	62.89
MRMR	62.3 $\pm$ 0.85	64.23
RFE	64.6 $\pm$ 0.84	62.80
Softmax	**73.5$\pm$ 0.78**	65.06
Sparsemax	72.6 $\pm$ 0.79	**65.39**

Conclusion

How do we select the self-supervised pretext tasks optimally towards solving a given downstream one ?

Conclusion

How do we select the self-supervised pretext tasks optimally towards solving a given downstream one ?

- Use Conditional Independence to predict the utility of a pretext-task towards solving a given downstream task.
- Extension to multi-task pretext task selection.
- Efficient way for SSL pretext-tasks exploration.

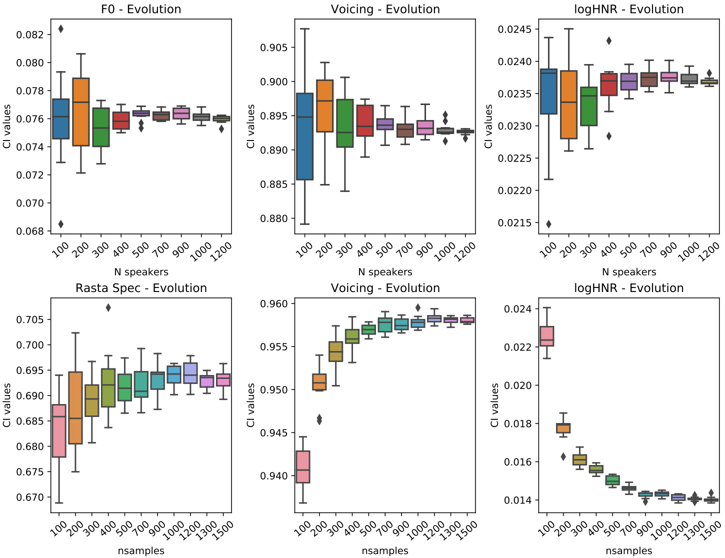
Thank you

- Thank you for your attention
- Open for questions

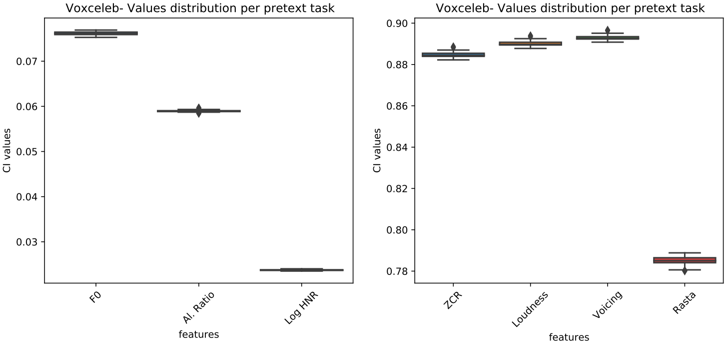
Changing the pretraining dataset

Table: Results observed retraining the Wav2vec2 model with and without weighted pretext tasks using the sparsemax method, on LibriSpeech 960. “Fr." and “Fine." also respectively refer to Frozen and Finetuned settings.

Selections	**LibriSpeech** *(WER % ↓)*	
	Fr.	Fine.
wav2vec 2.0 *BASE*	9.88	6.33
wav2vec 2.0 *BASE* + multitask SSL	**9.5**	**6.01**



Evolution of the CI estimation with different numbers of considered speakers for VoxCeleb (First row of plots) and number of samples for Medley (Second row of plots).



Boxplots of the CI values for every pretext tasks, when more than 200 speakers are considered. Voicing and Loudness are slightly overlapping, but otherwise, the values are separable. We divide the pretext-tasks in two groups according to their CI values for a better visualisation of the results.

Task change : Instrument Recognition

Table: Results observed with the proposed selection strategies on the two considered downstream instrument recognition tasks. Accuracy on the test set is computed for Medley-solos-DB while mean F1 Score is shown for OpenMIC. Higher is better for both.

Models	**Medley-solos** *(Acc% ↑)*	**OpenMIC-2018** *(mean-F1 ↑)*
PASE+ (Ravanelli, 2020)	None	64.1
Selections		
All	66.2 ± 0.83	62.89
MRMR	62.3 ± 0.85	64.23
RFE	64.6 ± 0.84	62.80
Softmax	**73.5± 0.78**	65.06
Sparsemax	72.6 ± 0.79	**65.39**
Sparsemax+	**76.1± 0.76**	66.0
Spectral+	74.6± 0.77	**67.7**