Pretext Tasks Selection for Multitask Self-Supervised Audio Representation Learning

Salah Zaiem[1] Titouan Parcollet[2] Slim Essid[1]

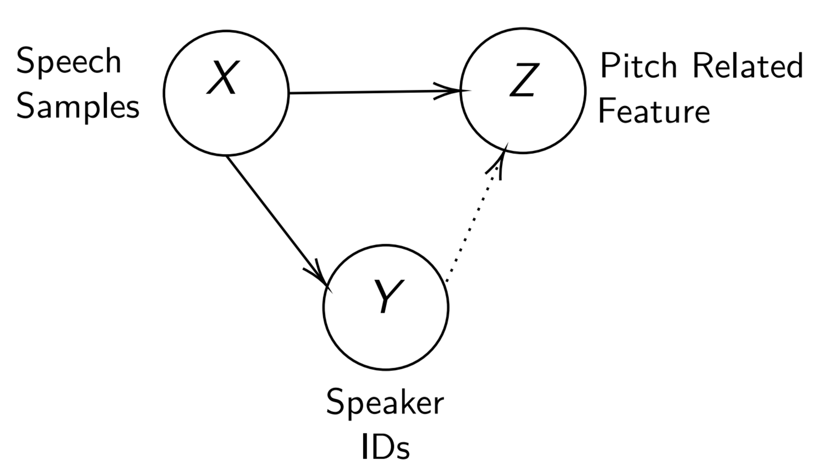
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France [2]LIA, Avignon Université

Abstract

Self-supervised allows leveraging unlabeled data to learn useful representations through solving pretext tasks. However, methods and common practices for combining such pretext tasks for better performance on the downstream task have not been explored and understood properly. We provide and **efficient and motivated** offline method allowing the **selection and weighting** of pretext tasks and validate it on audio data.

Main Idea

The more the pretext task labels are **independent** from the speech samples **conditionally on** the downstream labels, the better should the final performance be [2].



Individual Pretext task selection

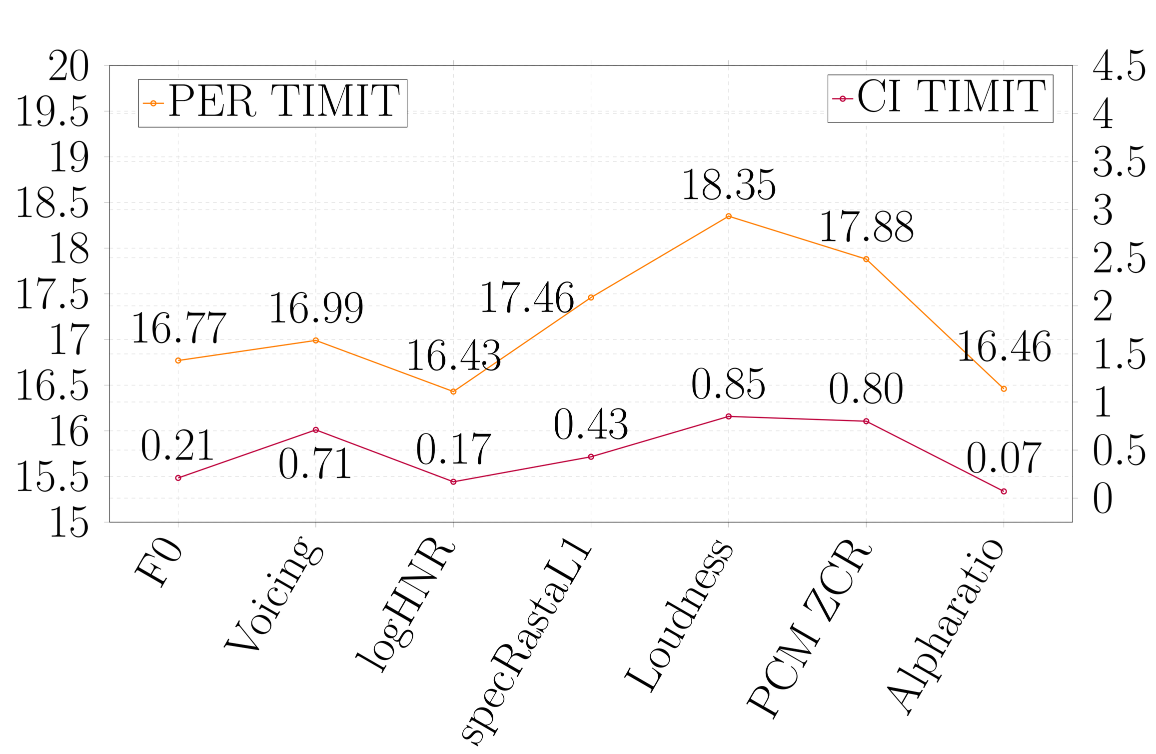
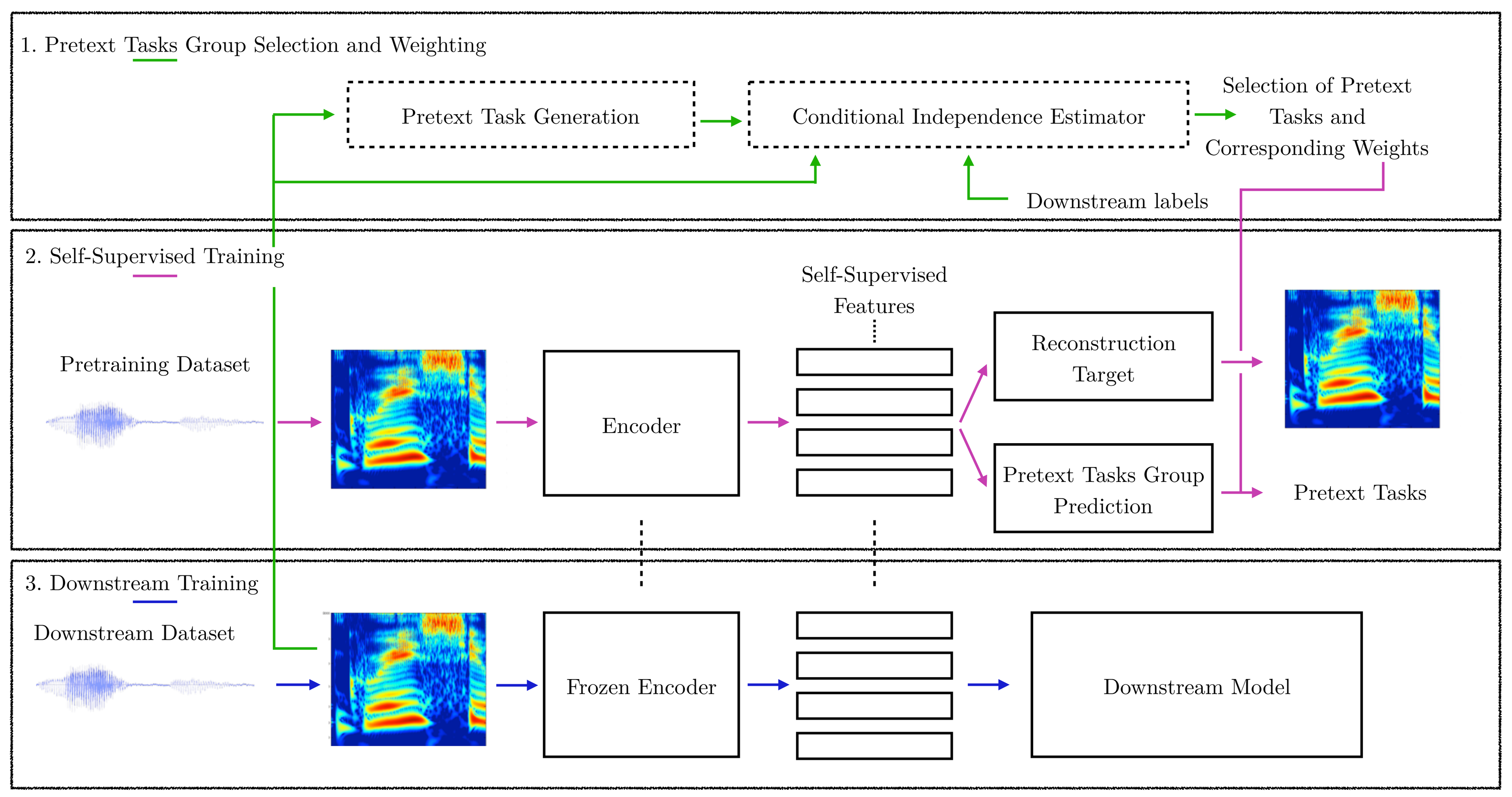


Figure 1:Phone Error Rate and CI estimate values on TIMIT for every considered pretext-task label [4]. We can observe the monotonic relation between the estimator and the downstream errors. Spearman correlation reaches 0.93 and Kendall τ : 0.81.

Multitask Selection and Weighting

Goal : given a set of k possible pretext-task labels $(Z_i)_{i\in[0,k]}$, we seek to estimate a set of parameters $(\lambda_i)_{i\in[0,k]}$ weighting the loss of every pretext-task label Z_i during the pre-training phase.

Loss function : $L_{SSL} = MSE_{mel} + MSE_{mfcc} + \Sigma_{i=1}^{k}\lambda_i \ell_1(Z_i)$



Constraints on the weights

- Positivity : $\lambda_i \geq 0 \Longrightarrow$ No adversarial learning.
- Constant sum : $\Sigma_i \lambda_i = 1$ enforced by $\lambda = softmax(W)$
- Sparsity to enforce selection, obtained with $\lambda = sparsemax(W)$

Respecting these constraints, the selected $(\lambda_i)_{i\in[0,k]}$ are the ones minimizing the CI criterion.

Results

Results observed with the proposed selection strategies on the two considered downstream tasks.

Models	**LibriSpeech** *(WER % ↓)* *No LM*	*LM*	**VoxCeleb1** *(EER % ↓)*	**IEMOCAP** *(Acc % ↑)*
PASE+ [3]	25.11	16.62	11.61	57.86
vq-wav2vec	17.71	12.80	10.38	58.24
Selections				
All	21.98 ± 0.36	11.70 ± 0.27	11.90± 0.32	56.4 ± 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

Results observed retraining the Wav2vec2 [1] model with and without weighted pretext tasks. The loss function for the third line here is : $L_{SSL} = L_{W2V} + \Sigma_{i=1}^{k}\lambda_i \ell_1(Z_i)$.

Selections	**LibriSpeech** *(WER % ↓)* Fr.	Fine.	**VoxCeleb1** *(EER % ↓)* Fr.	Fine.	**IEMOCAP** *(Acc % ↑)* 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**

Extension to Music

Table 1:Accuracy on the test set is computed for Medley-solos-DB while mean F1 Score is shown for OpenMIC. Higher is better for both. Sparsemax + is an experiment with additional pretext tasks in the starting pool. Spectral is an experiment replacing MFCCs with a combination of spectral representations.

Models	**Med. solos** *(Acc% ↑)*	**Op.MIC** *(mean-F1 ↑)*
PASE+	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**

Pretraining Datasets

- CommonVoice English (6.0 700 hours) and Audioset Musical instruments part (155 hours), respectively for speech and music pretraining.

Take-away messages

- Efficient way to select and weight pretext tasks depending on the downstream task of interest.
- Robust to downstream task, data type and pretraining dataset changes.
- Improves Sota methods on three considered downstream tasks.
- Code is available on github and within the SpeechBrain library for replication and further investigations.

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.
[2] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning, 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. Conditional independence for pretext task selection in self-supervised speech representation learning, 2021.