Self-supervised Speaker Verification with Adaptive Threshold and Hierarchical Training

Zehua Zhou, Haoyuan Yang, Takahiro Shinozaki
Tokyo Institute of Technology, Japan

Introduction

Problem

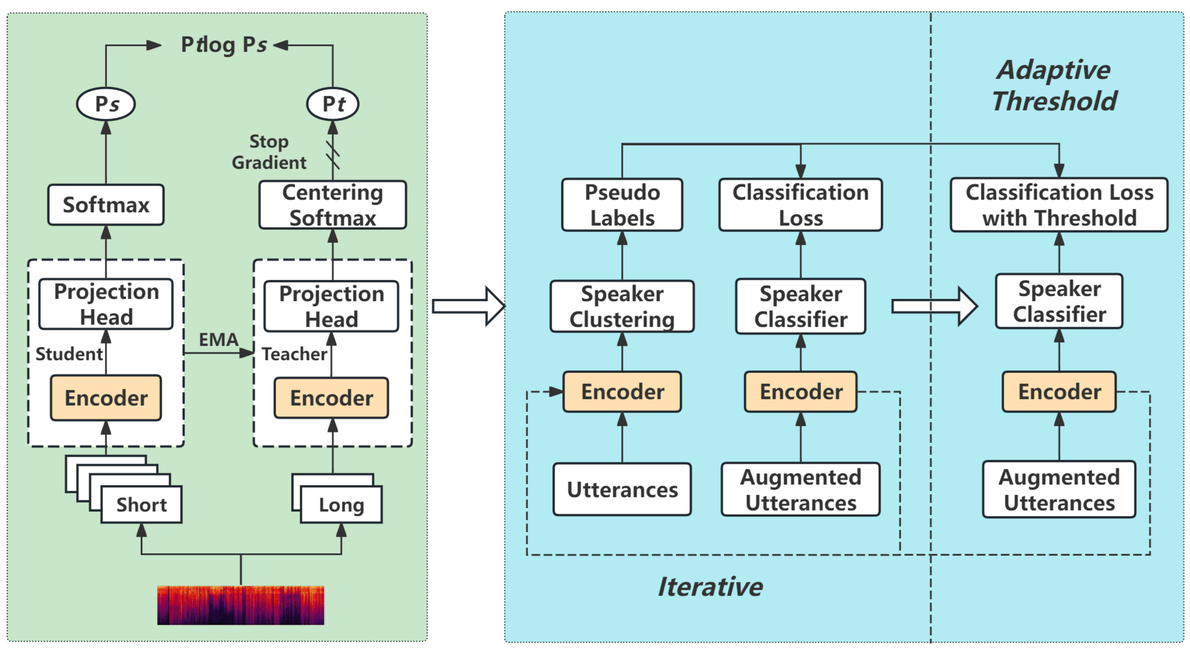
Current popular self-supervised speaker verification (SV) systems usually follow an iterative framework.

However, pseudo-labels generated by clustering algorithm are not always reliable and can degrade the model's performance during training.

Aim

To explore methods to reduce the negative impact of unreliable pseudo-labels and thereby improve the model's Equal Error Rate (EER) performance.

Related Work



Stage Ⅰ: DINO *Stage Ⅱ: Iterative self-supervised learning*

Stage1: Train the speaker encoder using the DINO-based loss to obtain speech representations.
Stage2: Utilize the clustering algorithm to generate pseudo-labels for each utterance. Then, iteratively train a new model in a self-supervised manner with the estimated pseudo-labels.

In such an iterative framework, how to select high-quality pseudo-labels has become the key. Study [1] sets a fixed threshold to differentiate between reliable and unreliable pseudo-labels, which lacks flexibility. Study [2] observed that clean data and noisy data exhibit some separation in the loss distribution. Study [3] in computer vision found that the latter layers of DNNs are much more sensitive to label noise, while their former counterparts are quite robust.

Proposed Methods

Adaptive Threshold

In stage 2, we propose using a Gaussian Mixture Model (GMM) with two components to fit the distribution of reliable and unreliable samples, expressed as $p(x) = \lambda_1 N(\mu_1, \sigma_1^2) + \lambda_2 N(\mu_2, \sigma_2^2)$.

The threshold τ is automatically determined to equalize the probabilities of two components: $p_1(\tau) = p_2(\tau)$, where $p_1(x) = \lambda_1 N(\mu_1, \sigma_1^2)$ and $p_2(x) = \lambda_2 N(\mu_2, \sigma_2^2)$. Then threshold τ will be applied to the AAM-Softmax. This method eliminates the need for manual threshold setting, addressing issues of inflexibility and potential inaccuracy.

Hierarchical Training

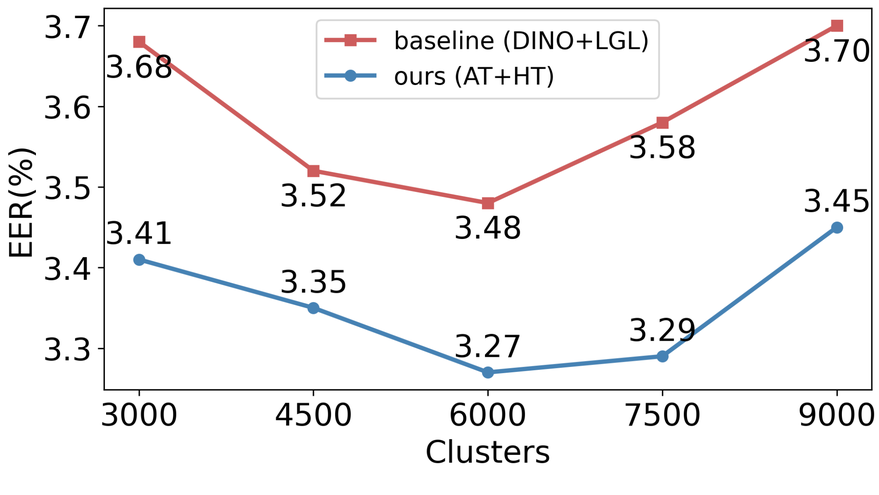
We partition the network structure in stage 2 into different parts and then apply hierarchical training. Specifically, rather than training the entire model at one time, we train the earlier layers in the network with more epochs, and then fix the previous layer and use fewer epochs to train the subsequent layers, countering the impact of unreliable pseudo-labels.

Dataset

Training: VoxCeleb2 dev set (1,092,009 utterances, 5,994 speakers)
Evaluation: Vox-O test set (4,874 utterances from 40 speakers)

Experimental Results

Type	Method	EER (%)
Contrastive Learning	Cai et al. [10]	3.45
	Tao et al. [13]	3.09
	Thienpondt et al. [29]	2.10
Non-contrastive Learning	DINO [18]	2.03
	+ LGL [13]	1.52
	+ AT	1.42
	+ + HT	**1.35**



- The adaptive threshold strategy results in a 6.6% improvement compared to the LGL strategy, which uses a fixed threshold.
- Futhermore, introducing the hierarchical training strategy can further enhance this improvement to 8.9%.
- Our methods can bring obvious and stable improvements across all given number of clusters.

Conclusion

- We proposed two strategies within the DINO-based self-supervised SV framework. Experiments showed their robustness and effectiveness in improving the model's performance on EER.
- Exploring hierarchical training at a finer granularity will be the future work.

References

[1] Tao et al., ICASSP, 6142–6146, 2022.
[2] Han et al., Interspeech, 4780–4784, 2022.
[3] Bai et al., NeurIPS, 24392–24403, 2021.