SJTU-AISPEECH System for VoxCeleb Speaker Recognition Challenge 2022

Zhengyang Chen[1,*], *Bing Han*[1,*], *Xu Xiang*[2,*], *Houjun Huang*[2,*], *Bei Liu*[1], *Yanmin Qian*[1]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
Department of Computer Science and Engineering, X-LANCE Lab, Shanghai Jiao Tong University
[2]AISpeech Ltd, Suzhou China

SJTU Cross Media
Language Intelligence Lab
上海交通大學跨媒體語言智能實驗室

AISPEECH 思必驰
专 注 人 性 化 的 智 能 语 音

Outline

- Track1 System Description (3rd place)
  - Data Processing
  - System Training & Evaluation
  - Models & Results
- Track3 System Description (4th place)
  - Domain Adaptation Strategy
    - Statistic Adaptation
    - Jointly Training Based Domain Adaptation
      - Dynamic Loss-gate and Label Correction
  - System Evaluation
  - Results

- Data Processing
  - Data augmentation
    - Speed perturbation
    - Additive noise and reverberation.
  - Acoustic feature
    - 80-dimensional Fbank

System Training & Evaluation

- Stage I: Pre-training
  - Optimizer: SGD
  - Training Objective:
    - AAM loss (margin 0.2) + K-subcenter + Inter-Topk [1]
  - Training segment length: 2s

- Stage II: Large Margin Fine-tuning
  - Optimizer: SGD
  - Training Objective:
    - AAM loss (margin 0.5) + K-subcenter
  - Training segment length: 6s

- Evaluation:
  - Cosine similarity
  - Adaptive score normalization
    - Estimate imposter cohorts from voxceleb2 dev set
  - Quality-aware score calibration
    - Simulate a trial from voxceleb2 dev set

[1] Zhao, Miao, et al. "Multi-Query Multi-Head Attention Pooling and Inter-Topk Penalty for Speaker Verification." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

Track1 System Description

Models & Results

Table 6: ***System results on track1 validation set and test set.*** *Without specific notation, the results are evaluated on the validation set.*

Index	Backbone	Pooling Method	Param #	EER (%)	minDCF
Online System					
S1	ResNet34-c64	MQMHA	27.8M	2.001	0.1375
S2	ResNet101-c32	MQMHA	23.8M	1.551	0.1029
S3	ResNet101-c64	MQMHA	68.7M	1.566	0.1034
S4	ResNet152-c32	MQMHA	27.7M	1.457	0.1036
S5	ResNet221-c32	MQMHA	31.6M	**1.420**	0.0976
Offline System					
S6	ResNet152-c32	Statistic Pooling	19.7M	1.609	0.0980
S7	Res2Net101-c32	Statistic Pooling	28.6M	1.480	**0.0890**
Fusion	-	-	-	1.056	0.0686
Fusion (test set)	-	-	-	1.911	0.1010

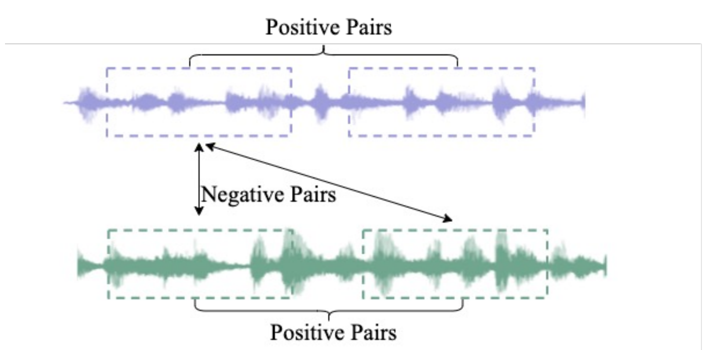
[1] Zhao, Miao, et al. "Multi-Query Multi-Head Attention Pooling and Inter-Topk Penalty for Speaker Verification." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

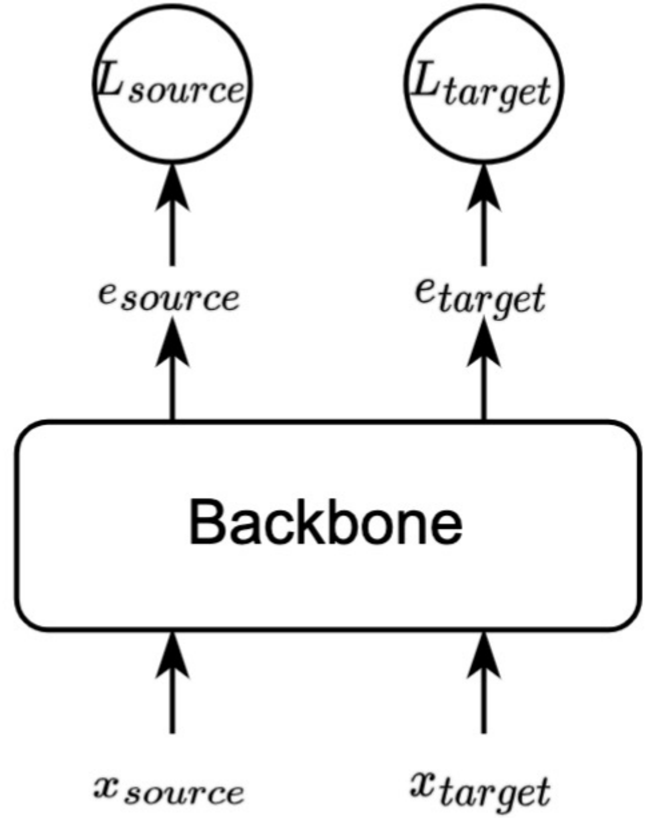
- Statistic Adaptation
  - Using statistics from track3 unsupervised data to mean-normalize the evaluation speaker embedding

Track3 System Description

Domain Adaptation Strategy

- Statistic Adaptation
- Jointly Fine-tuning Based Domain Adaptation
  - Source domain objective is the same as the pre-training system
  - Target domain objective
    - Self-supervised learning based angular prototypical loss (APL)
    - Classification loss based on estimated pseudo label
      - Two head classification loss (TCL)
      - One head classification loss (OCL)





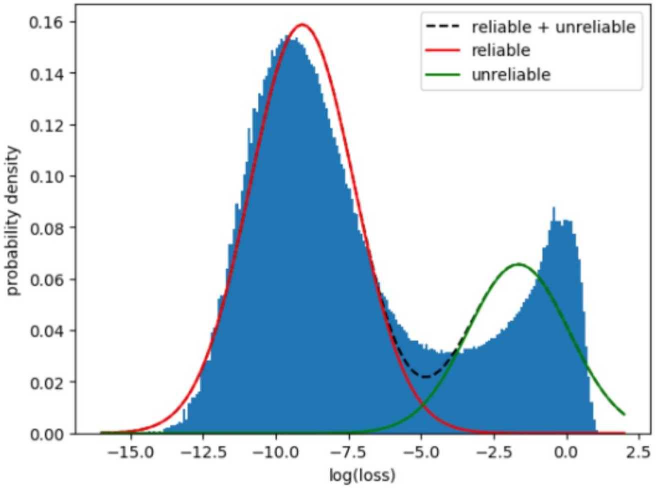
$$\mathcal{L} = \mathcal{L}_{source} + \mathcal{L}_{target}$$

Track3 System Description

Domain Adaptation Strategy

- Statistic Adaptation
- Jointly Training Based Domain Adaptation
  - Source domain objective is the same as the pre-training system
  - Target domain objective
    - Self-supervised learning based angular prototypical loss
    - Two head classification loss based on estimated pseudo labels
    - One head classification loss based on estimated pseudo labels

- Dynamic Loss-gate and Label Correction (DLG-LC) [1]



[1] B. Han, Z. Chen, and Y. Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," arXiv preprint arXiv:2208.01928, 2022.

- System Evaluation
  - Cosine similarity
  - Adaptive score normalization
    - Estimate imposter cohorts from track3 unsupervised set with pseudo labels
  - Quality-aware score calibration
    - Construct a trial from track3 labeled data

Ablation study on back-end processing method

Back-end Processing Method	EER (%)	minDCF
ResNet34-c64		
cosine scoring	14.60	0.5302
+ statistic adaptation	11.65	0.4552
++ as-norm	11.58 (11.735)	0.4032 (0.3959)
+++ score-calibration	9.950	0.4290

(): the results in bracket show the as-norm results when using ground-truth speaker labels

Comparison between different adaptation method

Adaptation Method	Backbone	EER (%)	minDCF
Statistic Adaptation	ResNet34-c64	9.950	0.4290
APL	ResNet34-c64	9.050	0.4063
TCL	ResNet34-c64	9.320	0.4254
TCL + DLG-LC	ResNet34-c64	9.060	0.4180
OCL	ResNet34-c64	8.825	0.4104
	ResNet101-c32	8.255	0.3771
	ResNet152-c32	8.095	0.3696
	ResNet101-c64	8.025	0.3670
	ResNet221-c32	8.255	0.3674
OCL + DLG-LC	ResNet152-c32	7.855	0.3681
Fusion	-	7.135	0.3290
Fusion (test set)	-	8.087	0.4370

:Pseudo labels are estimated from the APL system

APL: angular prototypical loss

TCL: two head classification loss

OCL: one head classification loss

Thanks!