The HCCL System for Semi-Supervised Domain Adaptation task of VoxSRC22

Zhuo Li*, Zhenduo Zhao*, Wenchao Wang, Pengyuan Zhang

Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

September 22, 2022

Overview

- Pseudo-labeling framework
  - Base model training with source labeled data
  - Embedding domain adaptation
  - Pseudo label generation
  - Model training with labeled source domain data and pseudo-labeled target domain data
  - Pseudo-label correction and retraining
- Supervised learning and self-Supervised learning

Base model training & Adaptation

- Base model training
  - Using models with as much variance as possible, either in terms of model structure or the training Protocol.

mdl	loss	vox2-train	t3-dev-EER ini	t3-dev-EER adapt
se-resnet34-32	circle	clean-fb64-sgd	16.86	14.29
cotnet	circle	clean-fb64-sgd	16.65	14.55
conformer	circle	aug-fb80-adam	16.95	14.14
ecapa-large	circle	aug-fb80-adam	18.02	14.62
se-resnet101-32	circle	aug-fb80-adam	14.06	11.90

- Adaptation
  - Aligning statistics between different domains
  - Aligning domain centers is easy and efficient, but aligning the variances need backends (LDA & PLDA)
  - We will explore variances alignment systematically in the future

Pseudo label generation

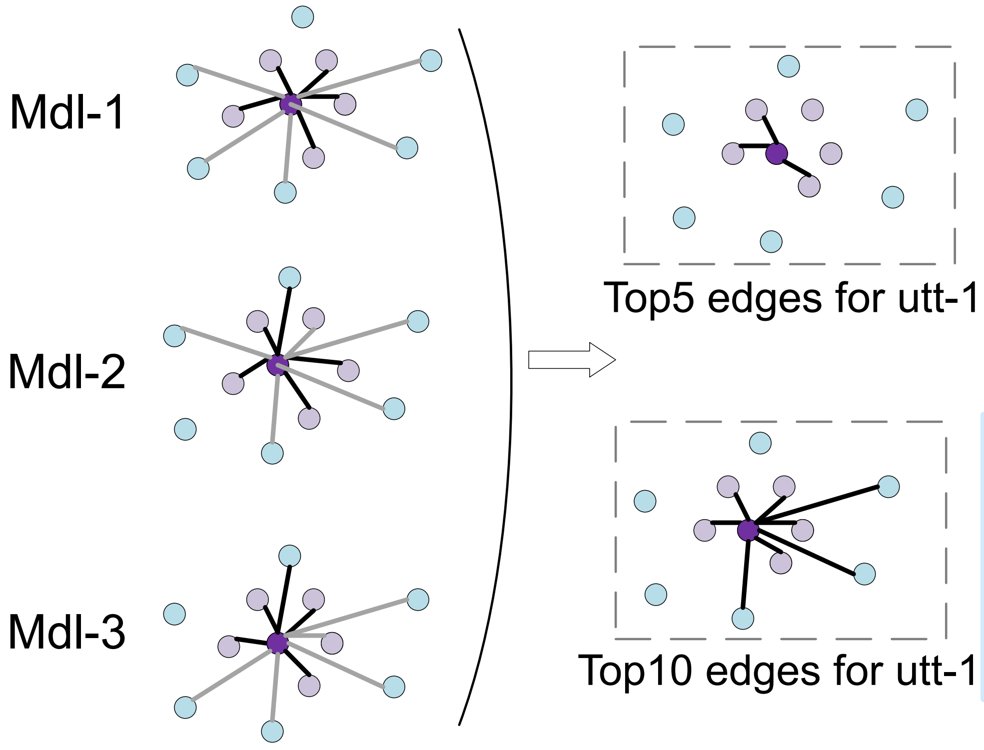
- Clustering algorithm :

  (a **progressive sub-graph** clustering algorithm based on **two Gaussian fitting** and **multi-model voting**)

- Key points:
  - finding high-confidence positive trials using a multi-model voting strategy based on the KNN affinity graph
  - utilizing connected sub-graphs to obtain pseudo labels
  - using iterative top-k information to gradually combine sub-classes
  - two Gaussian distributions fitting the intra-class score distribution to check for high-confidence edges

Pseudo label generation

High-confidence edges:

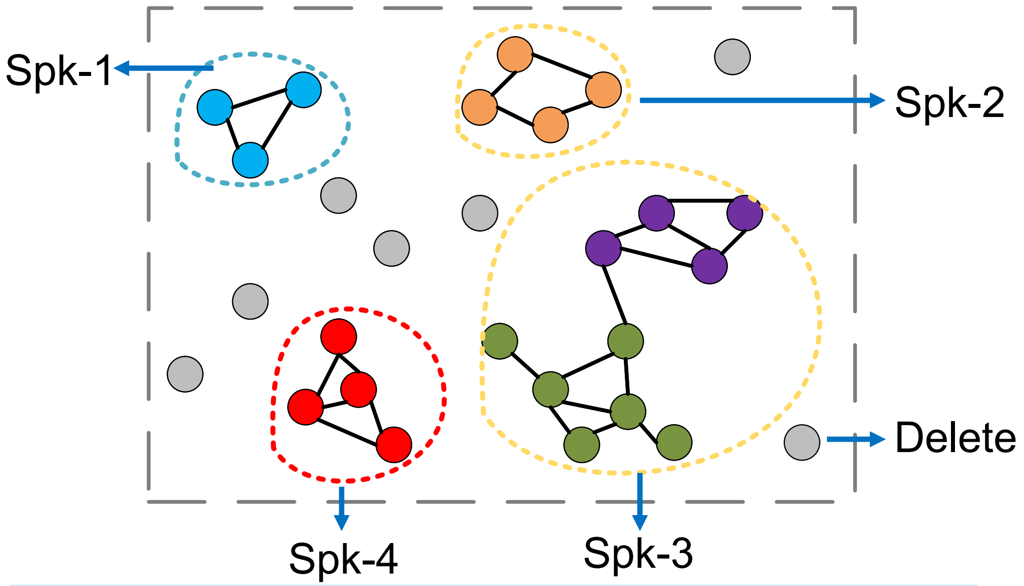


Notes1:
Whether or not to preserve edges also depends on similarity

Notes2:
Voting strategy can greatly decrease false positive rate

Constructing k-nearest neighbors graphs for utt-1 by voting

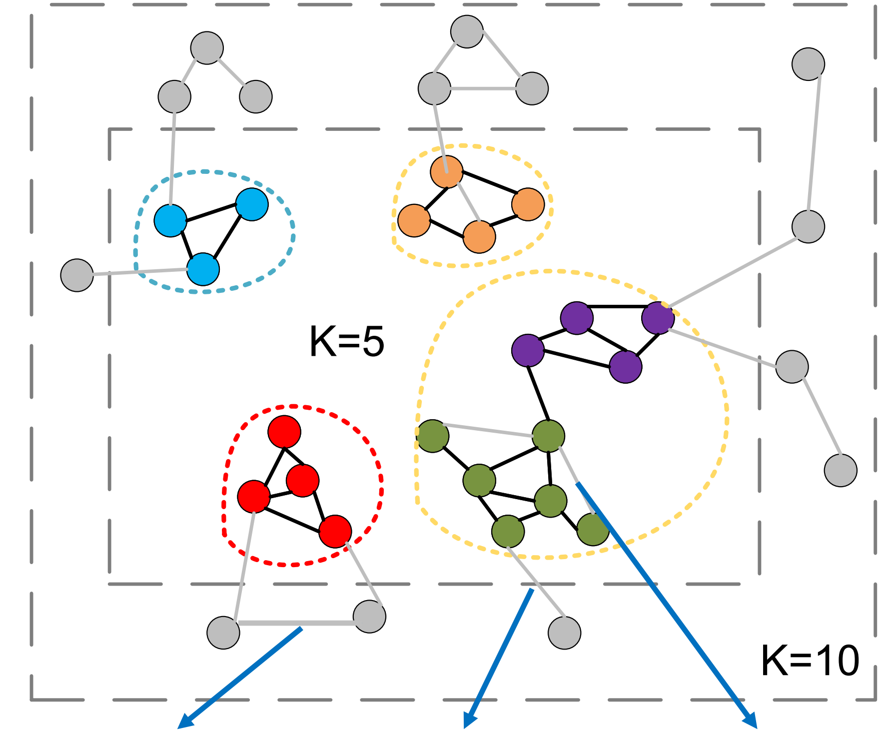
Connected sub-graphs:



Notes3:
Using connected sub-graph can greatly increase the intra-class diversity

Pseudo label generation

☐ Progressive :

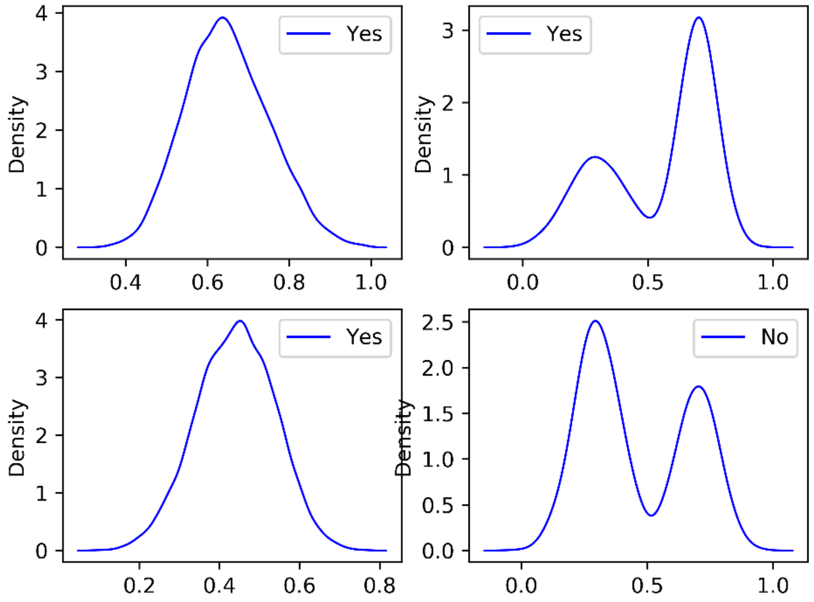


Edges_{new.new} Edges_{new.old} Edges_{old.old}

Notes1:
Preventing disastrous classes caused by error edges

☐ Two Gaussian fitting:

Cal similarity between all utts if combine



Notes1:
proper noisy label is okay

$\mu_1, \sigma_1, w_1; \mu_2, \sigma_2, w_2;$
represents parameters
max and min gaussian

$\mu_2 > th_{nm}$ *OR*
$w_1 > 0.5$ *OR*
$\mu_1 - \sigma_1 < (\ \mu_2 + \sigma_2) + \epsilon$

Model training with labeled source domain data and pseudo-labeled target domain data

- Stage 1
  - Subcenter is extremely important
  - Speed perturbation augmentation is used in all data
  - Both train the model from scratch and utilize models from Track1 as the pre-trained model are okay
  - For the latter, freezing the extractor in the beginning to make models be converged is necessary.
- Stage 2
  - CN-Celeb data without speed perturbation is used to finetune.
  - The VoxCeleb weights of the classification layer are preserved to prevent overfitting.
  - Expand chunksize to 6s and slightly increase constraint is effective

Pseudo-label correction and retraining

Error labels:

- 1label-vs-multispk (noisy labels)
- multilabel-vs-1spk (multi labels)

noisy labels:

- Subcenter is enough to correct

multi labels:

- Cal similarity of all audio in CN-Celeb to the two most similar class centers; denoted as s_i^1 and s_i^2;
- Split audios: s_i^1>0.5 and s_i^2<0.4 high-confidence; s_i^1>0.5 and s_i^2>0.4 median-confidence;
  s_i^1 < 0.5 low-confidence;
- Use audio with median-confidence to find ***multi labels***,
- Use the overlap between two labels to determine whether two labels need to be merged,
- Filter out audio that is low confidence, other audio is labeled by using predicted posterior probability.

Results and calibration

Table 4: *Results of systems for Track3. v0 means pseudo labels before correction, and v1 means after.*

	mdl	loss	train	t3-dev-EER	
				ini	calib
S1	Res2Net50	circle	v0-sgd	**8.45**	8.20
S2	ResNet34	circle	v0-sgd	8.61	**7.66**
S3	ECAPA-X4	circle	v0-sgd	9.57	8.81
S4	ECAPA-X4	circle	v0-adam	10.47	9.65
S5	ECAPA-X4	circle	v1-sgd	8.78	8.42

Table 5: *Results of systems that we submitted*

mdl	mode	dev-EER	eval-EER
S1	ini	8.45	8.07
S2	ini	8.61	8.64
S1+S2	ini	8.01	7.57
S1+S2+S3+S4	ini	7.87	7.40
S1+S2+S3+S4+S5	calib	**6.77**	**7.03**

Supervised&Self-Supervised Domain Adaptation

- Self-supervised learning requires no label
- Use supervised learning on labeled data, self-supervised learning on all data
- For supervised learning, we used circle loss; for self-supervised learning, we adopted DINO loss
- After pseudo-labeling, extract embeddings, averaged speaker-wise and appended after classification layer weights, continue training

Results

Table 3: *SSL-Assissted Domain Adaptation Results*

Data	Loss	Vox22-dev-t3	
		EER	$\mathrm{minDCF}_{0.05}$
Vox2dev	Circle	13.430	0.4848
Vox2dev+Track3	Circle+DINO	11.995	0.5160
Vox2dev+Track3-PL	Circle+DINO	10.055	0.4414

Reference

[1] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *conference of the international speech com- munication association*, 2018.

[2] H.Yamamoto,K.A.Lee,K.Okabe,andT.Koshinaka,“Speaker augmentation and bandwidth extension for deep speaker embed- ding,” *conference of the international speech communication as- sociation*, 2019.

[3] D.Snyder,G.Chen,andD.Povey,“Musan:Amusic,speech,and noise corpus,” *arXiv: Sound*, 2015.

[4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudan- pur, “A study on data augmentation of reverberant speech for ro- bust speech recognition,” *international conference on acoustics, speech, and signal processing*, 2017.

[5] M. Zhao, Y. Ma, M. Liu, and M. Xu, “The speakin system for voxceleb speaker recognition challange 2021,” *arXiv preprint arXiv:2109.01989*, 2021.

[6] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Ben- gio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.

[7] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large- scale speaker identification dataset,” *conference of the interna- tional speech communication association*, 2017.

[8] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, “Voxsrc 2020: The second voxceleb speaker recognition challenge,” *arXiv preprint arXiv:2012.06867*, 2020.

[9] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, “Voxsrc 2021: The third voxceleb speaker recognition challenge,” *arXiv preprint arXiv:2201.04583*, 2022.

[10] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Green- berg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chin- tala, V. Quenneville-Bélair, and Y. Shi, “Torchaudio: Build- ing blocks for audio and speech processing,” *arXiv preprint arXiv:2110.15018*, 2021.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *Proc. Interspeech 2020*, pp. 3830–3834, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] X.Ding,X.Zhang,N.Ma,J.Han,G.Ding,andJ.Sun,“Repvgg: Making vgg-style convnets great again,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog- nition*, 2021, pp. 13 733–13 742.

[14] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and- excitation networks,” *computer vision and pattern recognition*, 2018.

[15] Y. Shengyu, F. Xiang, Y. Jie, L. Jingdong, and P. Yiqian, “So- gou system for the voxceleb speaker recognition challenge 2021,” *arXiv: Sound*, 2021.

[16] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2net: A new multi-scale backbone architecture,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[17] Z. Li, “Explore long-range context feature for speaker verifi- cation,” *CoRR*, vol. abs/2112.07134, 2021. [Online]. Available: https://arxiv.org/abs/2112.07134

[18] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” *Proc. Interspeech 2018*, pp. 2252–2256, 2018.

[19] T. Stafylakis, J. Rohdin, and L. Burget, “Speaker embeddings by modeling channel-wise correlations,” in *Interspeech*, 2021, Con- ference Proceedings.

[20] R.Xiao,X.Miao,W.Wang,P.Zhang,B.Cai,andL.Luo,“Adap- tive Margin Circle Loss for Speaker Verification,” in *Proc. Inter- speech 2021*, 2021, pp. 4618–4622.

[21] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, “Sub-center ar- cface: Boosting face recognition by large-scale noisy web faces,” in*ECCV*. Springer,2020,ConferenceProceedings,pp.741–757.

[22] M. Zhao, Y. Ma, M. Liu, and M. Xu, “The speakin system for vox- celeb speaker recognition challange 2021,” *arXiv: Sound*, 2021.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chil- amkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Py- torch: An imperative style, high-performance deep learning li- brary,” *neural information processing systems*, 2019.

[24] L. N. Smith, “Cyclical learning rates for training neural net- works,” *workshop on applications of computer vision*, 2015.

[25] S.Cumani,P.D.Batzu,D.Colibro,C.Vair,P.Laface,andV.Vasi- lakakis, “Comparison of speaker recognition approaches for real applications,” *conference of the international speech communica- tion association*, 2011.

[26] J. Thienpondt, B. Desplanques, and K. Demuynck, “The idlab voxsrc-20 submission: Large margin fine-tuning and quality- aware score calibration in dnn based speaker verification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5814– 5818.

[27] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, “Cn-celeb: multi-genre speaker recogni- tion,” *Speech Communication*, vol. 137, pp. 77–91, 2022.

[28] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” *international conference on acous- tics, speech, and signal processing*, 2019.

[29] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, “Mfa-conformer: Multi-scale feature aggrega- tion conformer for automatic speaker verification,” *arXiv preprint arXiv:2203.15249*, 2022.

[30] Y. Li, T. Yao, Y. Pan, and T. Mei, “Contextual transformer net- works for visual recognition,” *IEEE Transactions on Pattern Anal- ysis and Machine Intelligence*, 2022.

[31] B.Sun,J.Feng,andK.Saenko,“Correlationalignmentforunsu- pervised domain adaptation,” in *Domain Adaptation in Computer Vision Applications*. Springer, 2017, pp. 153–171.

[32] K.A.Lee,Q.Wang,andT.Koshinaka,“Thecoral+algorithmfor *unsupervised domain adaptation of plda,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5821–5825.*

[33] R. Li, W. Zhang, and D. Chen, “The coral++ algorithm for un- supervised domain adaptation of speaker recognition,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7172–7176.

[34] S.Ioffe,“Probabilisticlineardiscriminantanalysis,”inEuropean Conference on Computer Vision. Springer, 2006, pp. 531–542.

[35] P. Kenny, “Bayesian speaker verification with, heavy tailed pri- ors,” Proc. Odyssey 2010, 2010.

[36] Chen Z , Wang S , Qian Y . Self-Supervised Learning Based Domain Adaptation for Robust Speaker Verification[J]. IEEE, 2021.

Thanks