Speech home

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

Robust Training for Speaker Verification against Noisy Labels

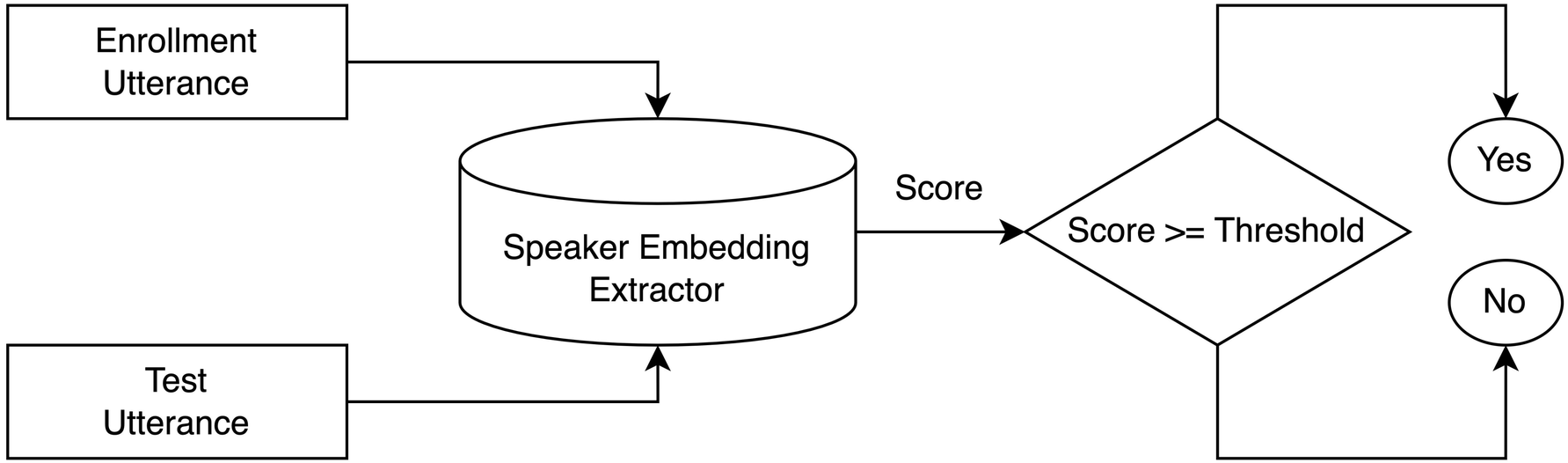
Zhihua Fang[1,2] Liang He[1,2,3,†] Hanhan Ma[1,2] Xiaochen Guo[1,2] Lin Li[4]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
[2] Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830017, China
[3] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[4] School of Electronic Science and Engineering, Xiamen University, Xiamen 361005, China

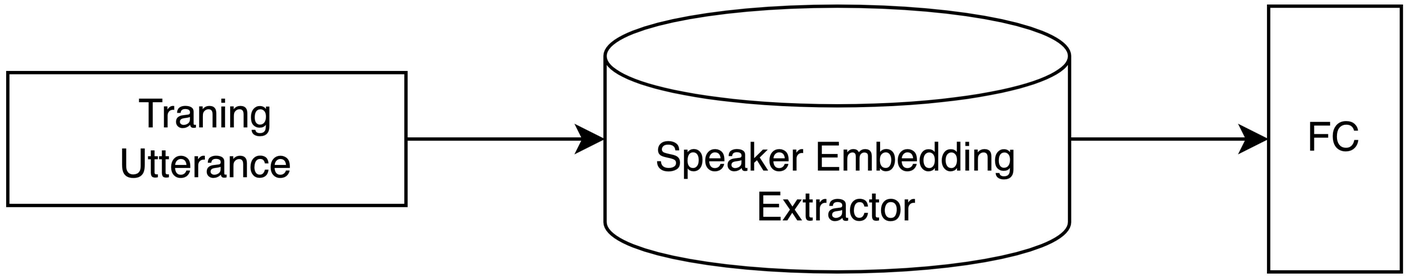
https://github.com/PunkMale/OR-Gate

Background | What is speaker verification?

Test:



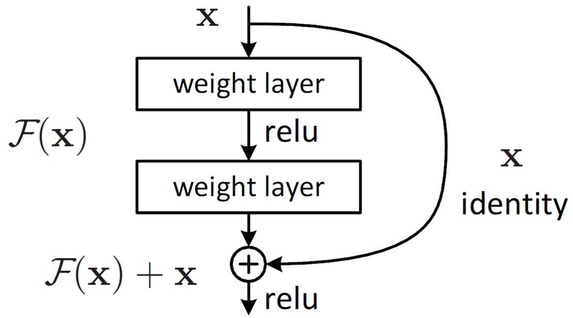
Train:



Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learn-ing: An overview," Neural Networks, vol. 140, pp. 65–99, 2021.

Background | Why DNNs work?

1. advanced deep network structures, such as ResNet;
2. powerful computer hardware, such as GeForce RTX 4090;
3. large-scale datasets with labeling, such as Voxceleb.



But noisy labels are inevitable!

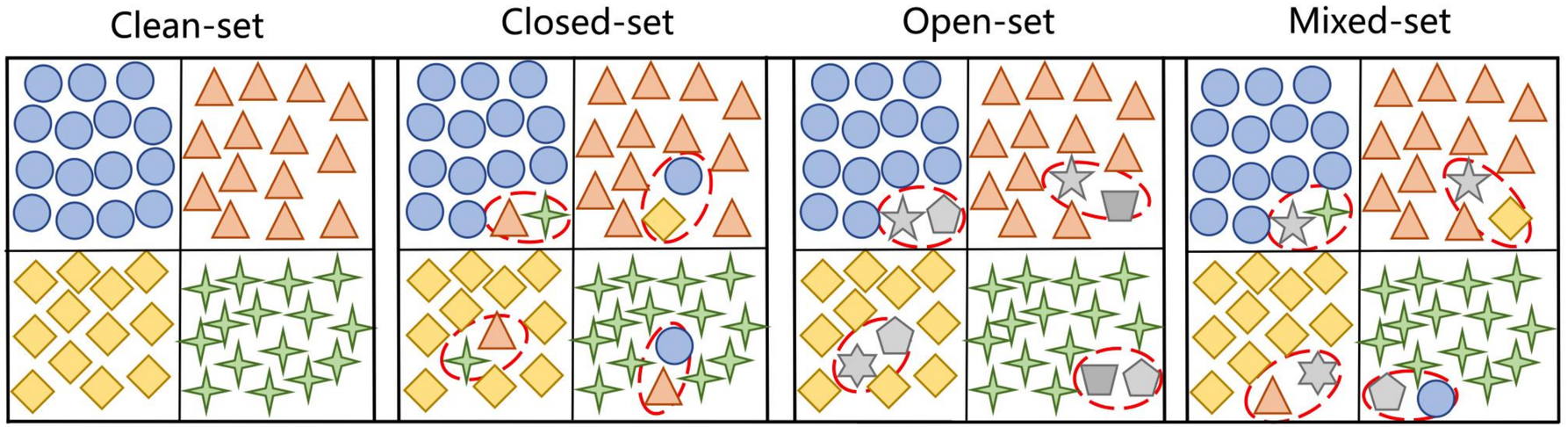
Background | What is noisy label?

◆What is noisy labels?

- A label that is incorrectly labeled in a dataset is a noisy label (e.g., it was originally speaker a, but was labeled as speaker b).

◆Types of noisy labels:

- Closed-set, Open-set, Mixed-set.

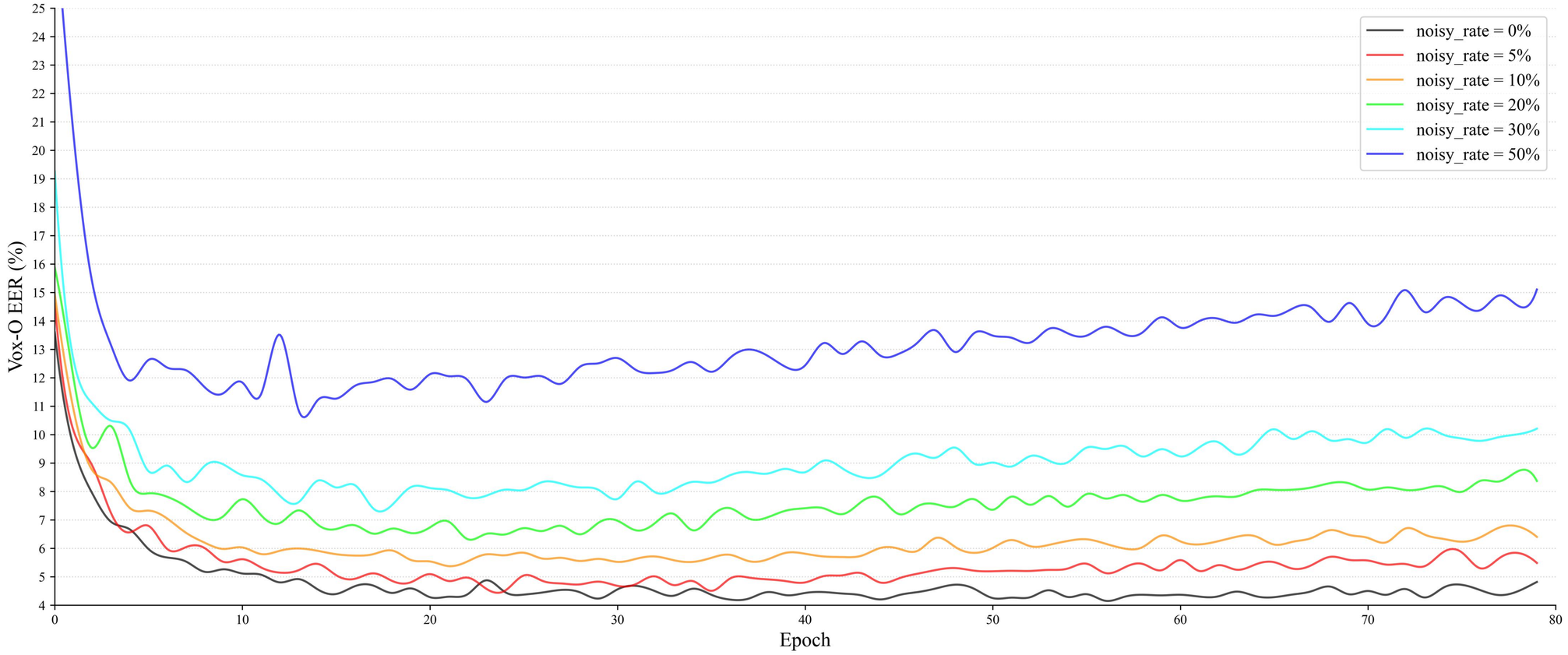


Lin Li, Fuchuan Tong, and Qingyang Hong, "When speaker recognition meets noisy labels: Optimizations for front-ends and back-ends," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1586–1599, 2022.

Background | The hazards of noise labels

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

◆ **The wrong label leads to the calculation of the wrong gradient descent direction, which ultimately leads to poor generalization of the model to clean test data .**

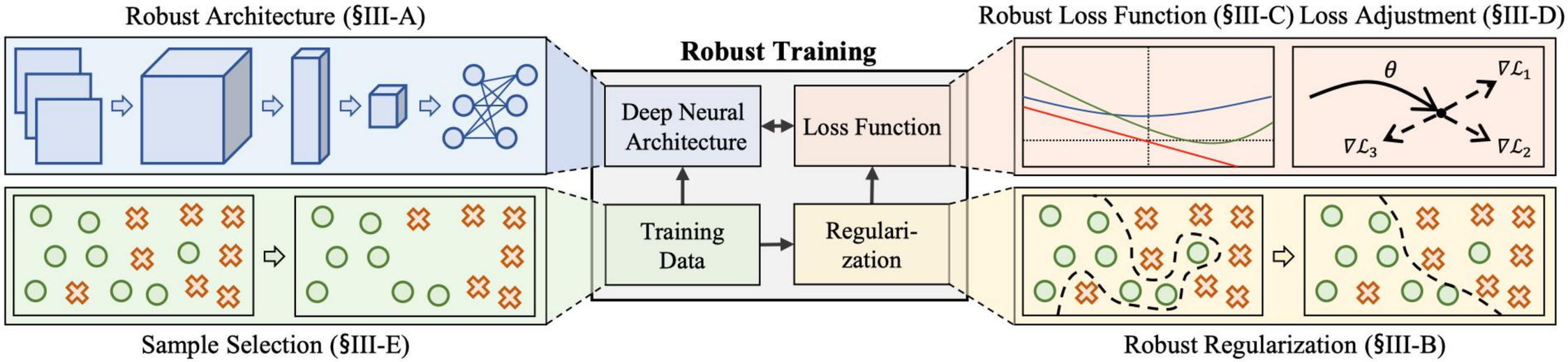


Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy, "The devil of face recognition is in the noise," in Proceedings of the European Conference on Computer Vision (ECCV), September 2018.

Related work | Learning with noisy labels

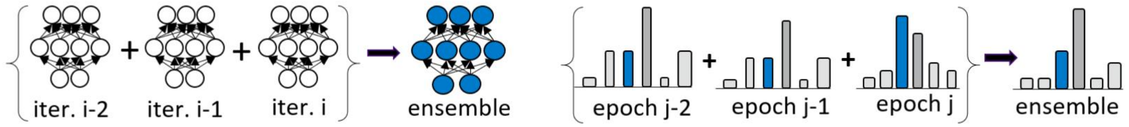
◆Noise label learning methods

- **Robust architecture:** usually a noise transfer layer is added after the model
- **Robust regularization:** prevents overfitting of the model to the sample by regularization methods
- **Robust loss function:** divided into Robust Loss Function and Loss Adjustment
- **Sample selection:** the network parameters are updated by selecting reliable sample labels through a well-designed network structure and training strategy



Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee, "Learning from noisy labels with deep neural networks: A survey," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–19, 2022.

Related work | *SELF (ICLR'20)*



(a) Model ensemble (Mean teacher)

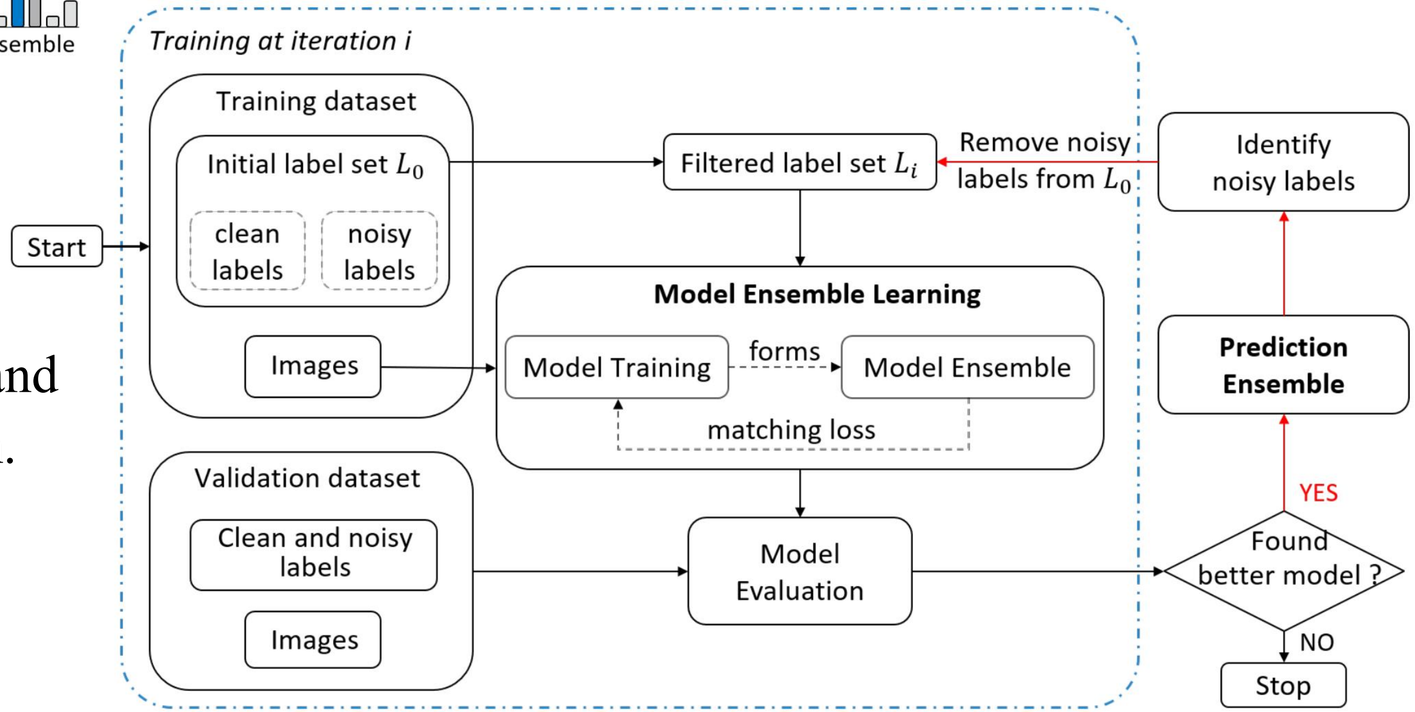
(b) Predictions ensemble

◆ Methodology

- Prediction ensemble
- a moving-average of the ensemble models and predictions to improve the filtering decision.

◆ Differences with speaker models

- Large variation in the number of classes
- Slow growth in speaker model prediction accuracy

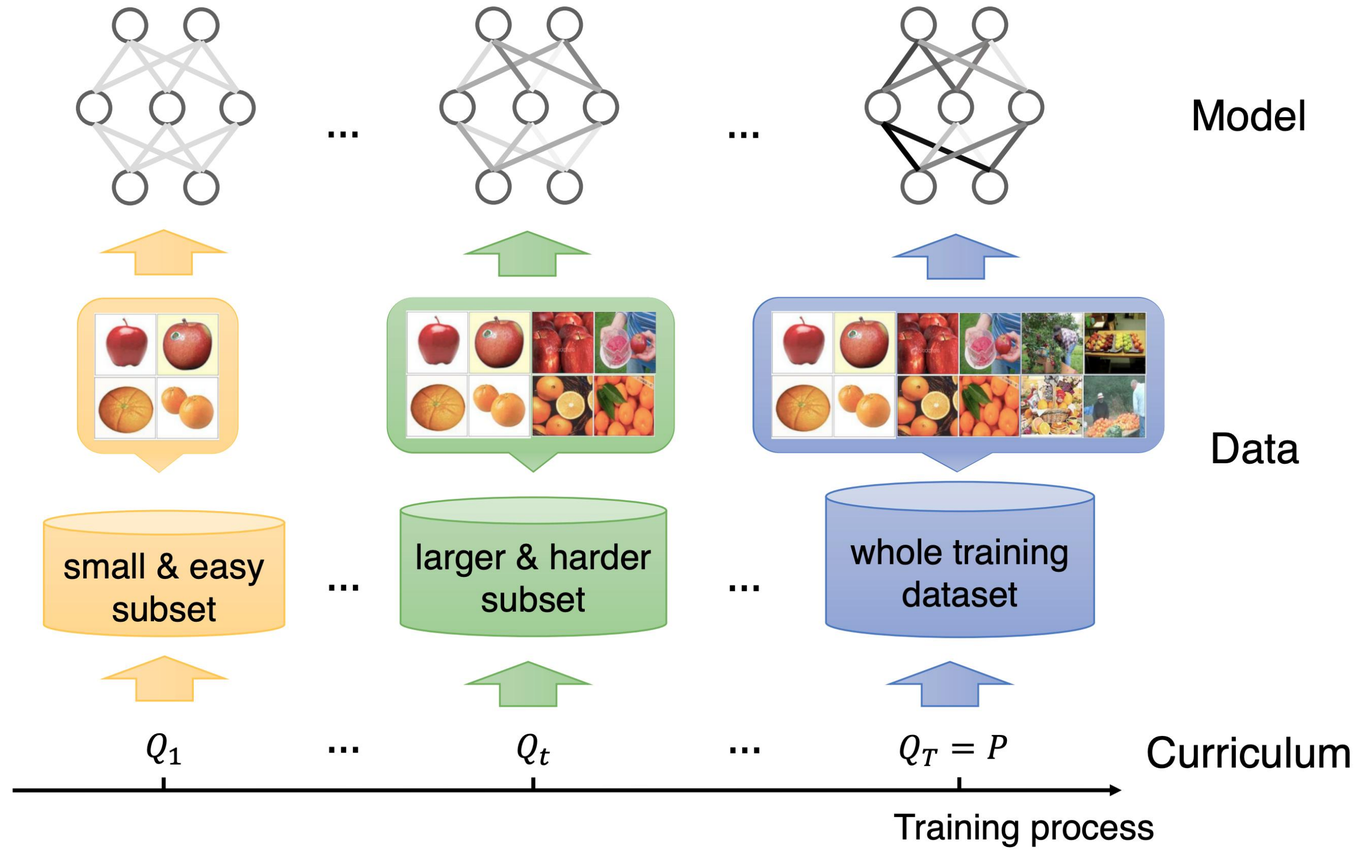


Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox, "Self: Learning to filter noisy labels with self-ensembling," in International Conference on Learning Representations, 2020.

Related work | Curriculum Learning

◆ **Definition: Let the model start learning from easy samples first and gradually progress to complex samples and knowledge.**

- **Predefined CL**
- **Automatic CL**
  1. Self-paced Learning
  2. Transfer Teacher
  3. RL Teacher
  4. Other Automatic CL



X. Wang, Y. Chen and W. Zhu, "A Survey on Curriculum Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, pp. 4555-4576.

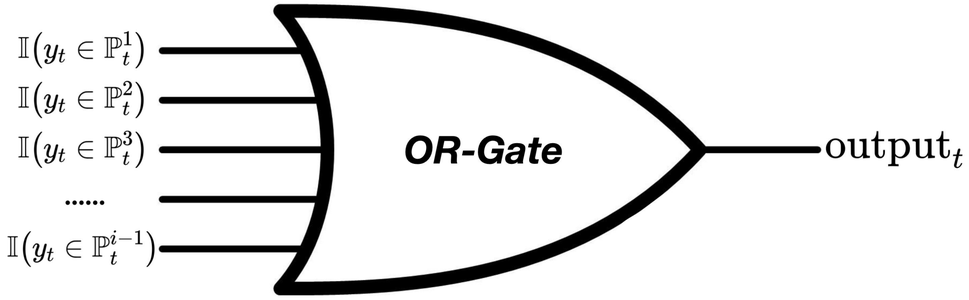
Methodology | Overview

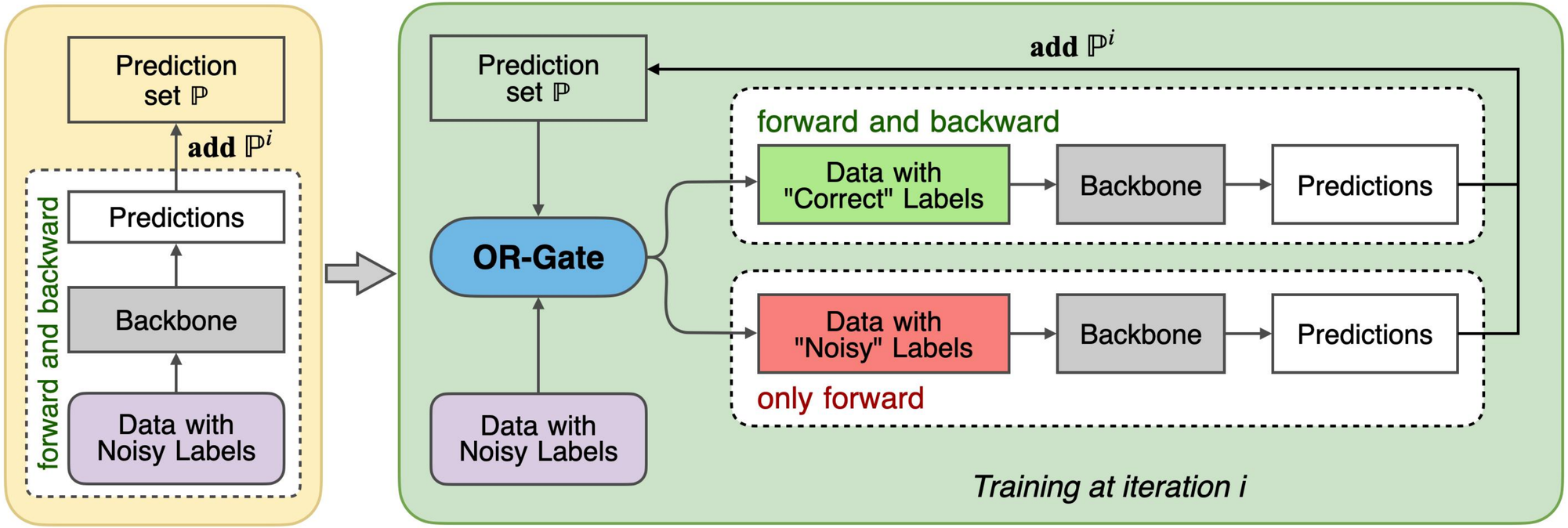
INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

◆ Two-Stage Learning

1. Early learning: Training with all data.
2. Self-confident learning: Use the ***OR-Gate*** to divide the data into reliable and unreliable data and process them separately.

* $\mathrm{P}^{i}{}_{t}$: The top k predictions for the t-th sample at the i-th epoch



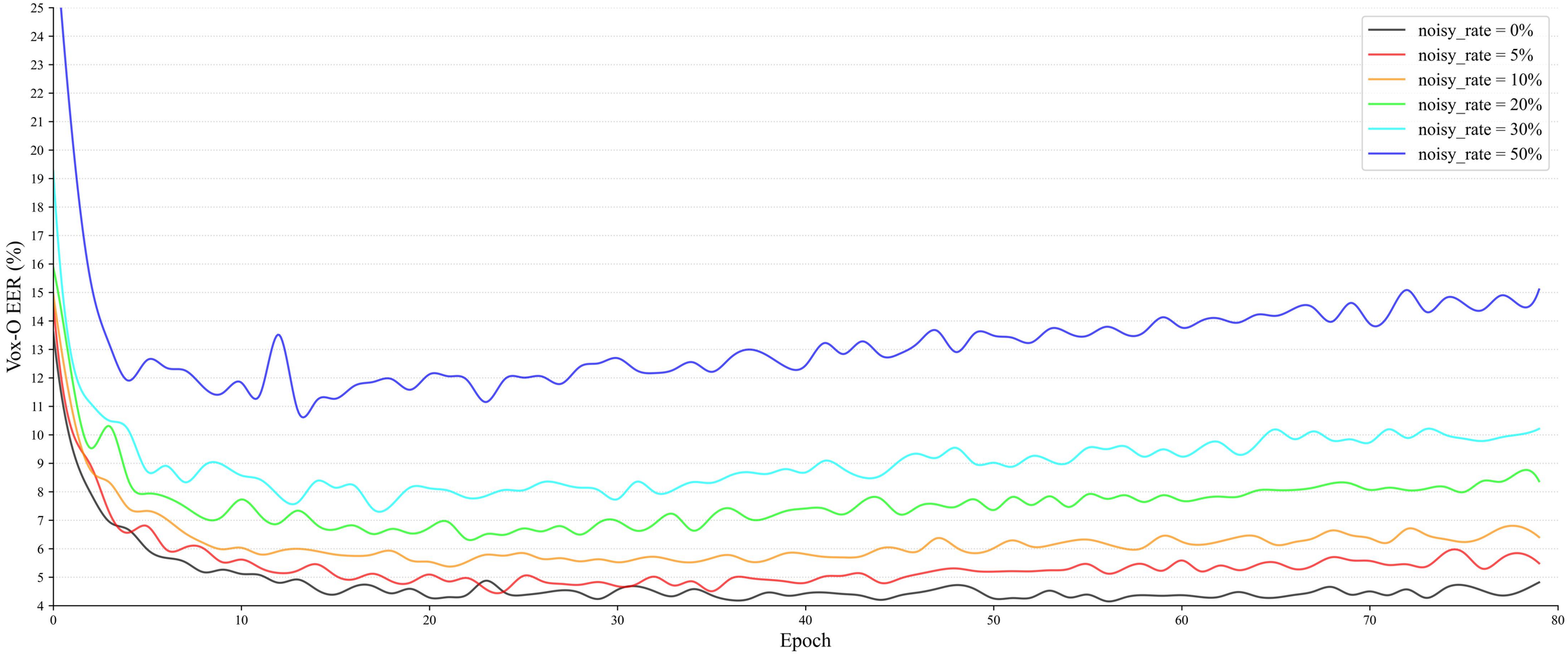


Stage I: Early learning

Stage II: Iterative self-confident learning

Methodology | Why Early Learning?

◆ DNN will fit the clean labels first, then the noisy labels



Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda, "Adaptive early-learning correction for segmentation from noisy annotations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 2606–2616.

Methodology | Why use the *OR-Gate* and top-*k* mechanism?

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

- **Fluctuations in the network parameters during training, and fluctuations in the network's predictions for different periods of the same sample;**
- **Hard samples are difficult to be accurately predicted in classification tasks;**
- **The key to how to balance the selection of hard and noisy samples is in the k-value.**

System		Proportion of Noisy Labels					
		0%	5%	10%	20%	30%	50%
Baseline		4.36	5.30	6.24	7.99	9.78	14.39
OR-Gate	$k = 1$	5.21	4.81	5.15	5.58	6.17	8.55
	$k = 5$	4.67	4.23	4.31	4.57	5.21	7.16
	$k = 10$	4.13	4.48	4.22	4.62	4.62	6.56
	$k = 20$	4.26	4.34	4.37	4.45	4.64	6.18
	$k = 30$	4.19	4.15	3.99	4.45	4.79	6.28
	$k = 40$	4.33	4.13	4.31	4.46	4.77	5.95
	$k = 50$	4.04	4.21	4.27	4.17	4.49	5.74
	$k = 70$	4.32	4.22	4.16	4.50	4.63	5.89
	$k = 90$	4.27	4.07	4.22	4.28	4.41	5.53

Results | EER(%) comparisons on VoxCeleb 1 & 2

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

- **The performance of the baseline deteriorates rapidly as the noise rate increases.**
- **Our method outperforms the baseline because our method follows the idea of curriculum learning.**
- **Our method outperforms other methods when training data with noisy labels.**

Table 1: *EER(%) comparisons on VoxCeleb1 with different proportions of noisy labels added.*

Test Data		Vox-O (original test pairs)					
η		0%	5%	10%	20%	30%	50%
Baseline		4.36	5.30	6.24	7.99	9.78	14.39
SELF [20] (Re-implemented)		5.52	5.94	5.90	6.34	7.50	11.90
LNCL+Sub-AM [22]	+ Cosine	4.64	4.83	5.11	5.45	5.56	6.32
	+ PLDA	3.92	4.45	5.02	5.47	5.96	7.29
	+ NL-PLDA	**3.92**	4.32	4.63	4.85	5.04	5.65
OR-Gate		4.08	**4.07**	**4.22**	**4.28**	**4.41**	**5.53**

Table 2: *EER(%) comparisons on VoxCeleb2 with different proportions of noisy labels added.*

Test Data		Vox-O (original test pairs)						Vox-H (hard test pairs)					
η		0%	5%	10%	20%	30%	50%	0%	5%	10%	20%	30%	50%
Baseline		1.69	1.72	1.90	2.21	2.88	4.32	**2.89**	3.11	3.25	3.64	4.42	6.74
LNCL+Sub-AM [22]	+ Cosine	**1.63**	**1.63**	1.64	1.69	1.80	2.25	3.02	3.05	3.06	3.26	3.26	3.77
	+ PLDA	1.71	1.78	1.81	2.05	2.15	2.86	3.03	3.10	3.15	3.57	3.76	4.76
	+ NL-PLDA	1.70	1.72	1.73	1.75	1.77	2.19	3.02	3.07	3.10	3.20	3.33	3.58
OR-Gate		1.64	1.65	**1.62**	**1.67**	**1.72**	**1.97**	2.93	**2.71**	**2.77**	**2.96**	**2.87**	**3.11**

Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proceedings of the 26th Annual International Conference on Machine Learning, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 4148.

Results | Precision and Recall of Divided Data

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

- **Precision: How many labels are correct for the selected data.**
- **Recall: How many of the correct labels have been selected.**

Table 3: *Precision and Recall of selecting clean labels at different periods of training.*

Metric	Precision						Recall					
η	0%	5%	10%	20%	30%	50%	0%	5%	10%	20%	30%	50%
VoxCeleb 1												
epoch_6	**1.0000**	**0.9999**	**0.9999**	**0.9999**	**0.9999**	**1.0000**	0.9523	0.9373	0.9161	0.8298	0.7018	0.3054
epoch_{10}	1.0000	0.9998	0.9997	0.9996	0.9992	0.9974	0.9857	0.9841	0.9800	0.9603	0.9250	0.6810
epoch_{20}	1.0000	0.9997	0.9995	0.9989	0.9980	0.9928	0.9942	0.9940	0.9933	0.9897	0.9780	0.8123
epoch_{50}	1.0000	0.9995	0.9990	0.9981	0.9963	0.9876	0.9975	0.9975	0.9971	0.9961	0.9916	0.8586
epoch_{80}	1.0000	0.9995	0.9988	0.9976	0.9953	0.9847	**0.9980**	**0.9982**	**0.9978**	**0.9969**	**0.9929**	**0.8598**
VoxCeleb 2												
epoch_5	**1.0000**	**0.9999**	**0.9999**	**0.9998**	**0.9999**	**0.9999**	0.9774	0.9743	0.9706	0.9545	0.9202	0.7093
epoch_{10}	1.0000	0.9999	0.9997	0.9994	0.9992	0.9979	0.9842	0.9837	0.9829	0.9806	0.9762	0.9369
epoch_{20}	1.0000	0.9998	0.9996	0.9991	0.9986	0.9961	0.9865	0.9862	0.9858	0.9845	0.9818	0.9499
epoch_{40}	1.0000	0.9997	0.9995	0.9988	0.9980	0.9945	0.9880	0.9878	0.9876	0.9866	0.9841	0.9552
epoch_{60}	1.0000	0.9997	0.9994	0.9986	0.9977	0.9937	**0.9888**	**0.9886**	**0.9884**	**0.9873**	**0.9850**	**0.9569**

Results | Ablation experiments

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

- **The performance of the model without early learning is much worse than the baseline, which indicates that it is only through early learning that a speaker model with basic recognition ability can be obtained for "self-confident" learning of Stage II.**
- **Adding the top-k mechanism to the *OR-Gate* can improve the performance more than not adding it, which indicates that the top-k mechanism can help the model select more hard samples to improve the classification ability of the model and get higher quality speaker embedding.**

Table 4: *EER (%) comparisons of the ablation experiments on VoxCeleb 1.*

Test Data	Vox-O (original test pairs)					
η	0%	5%	10%	20%	30%	50%
Baseline	4.36	5.30	6.24	7.99	9.78	14.39
OR-Gate w/o early learning	5.99	6.27	6.87	8.37	13.05	25.83
OR-Gate w/o top-k mechanism	5.21	4.81	5.15	5.58	6.17	8.55
OR-Gate	**4.08**	**4.07**	**4.22**	**4.28**	**4.41**	**5.53**

Thanks

https://arxiv.org/abs/2211.12080

https://github.com/PunkMale/OR-Gate

fangzhihua@stu.xju.edu.cn
heliang@mail.tsinghua.edu.cn

微信搜一搜
AuroraLab

* This slides will be uploaded to GitHub