The DKU-Tencent System for the VoxCeleb Speaker Recognition Challenge 2022

Xiaoyi Qin, Na Li, Yuke Lin, Yiwei Ding , Chao Weng, Dan Su, Ming Li

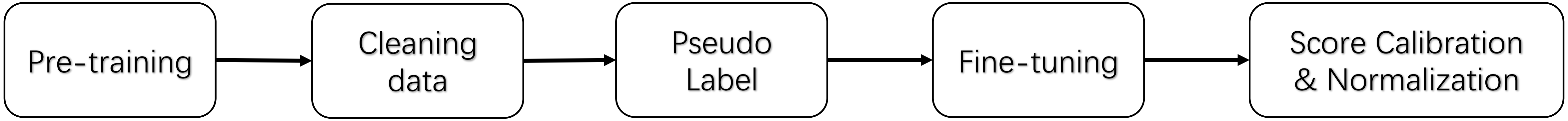
--Track 3

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Tencent AI Lab

Reporter：Xiaoyi Qin

Pipeline



Pre-training

Cleaning data

Pseudo Label

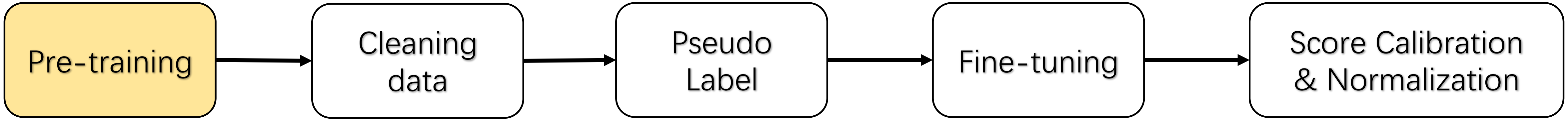
Fine-tuning

Score Calibration & Normalization

The training strategy following our FFSVC2022 task2 baseline system.

https://github.com/FFSVC/FFSVC2022_Baseline_System

1 Pre-training



Pre-training → Cleaning data → Pseudo Label → Fine-tuning → Score Calibration & Normalization

Training with Vox2Dev data：

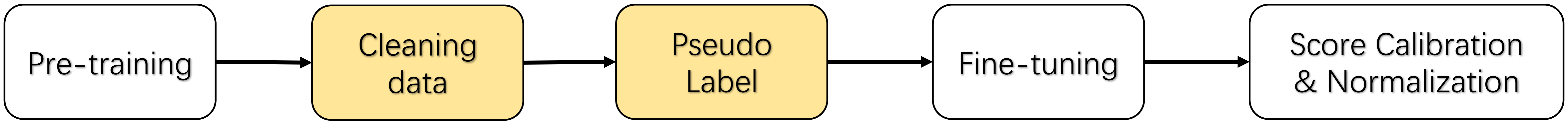
- SimAM-ResNet34-ASP
- ResNet101-ASP
- ResNet152-Stats
- ResNet152-ASP
- ResNet221-Stats
- RepVGG-ASP
- Res2Net101-Std
- SE-ResNet101-ASP

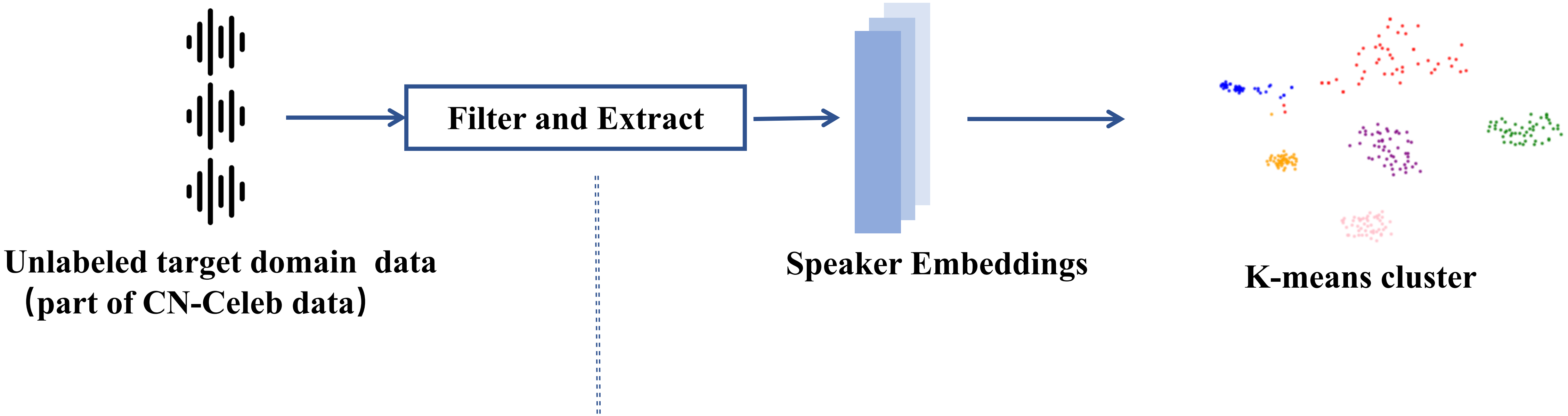
Data augmentation:

- An Online 3-fold speed perturbation is implemented (Spk Aug).
- On-the-fly data augmentation（Noise/RIR/Tempo/Vol）

Loss function ： ArcFace (m=0.2,s=32)

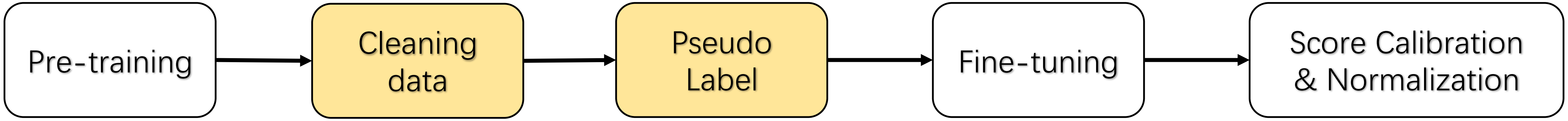
2 Cleaning data



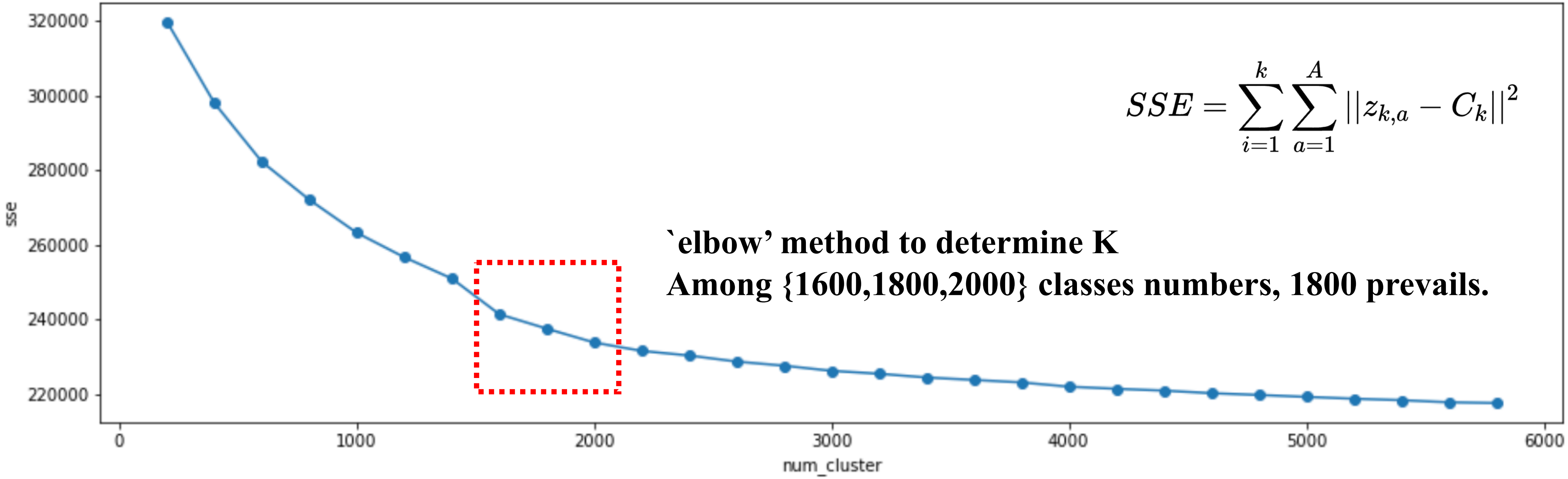


The audios whose duration is below 1s are removed.
(too short duration audio may not contain text information)

3 Pseudo label



Pre-training → Cleaning data → Pseudo Label → Fine-tuning → Score Calibration & Normalization



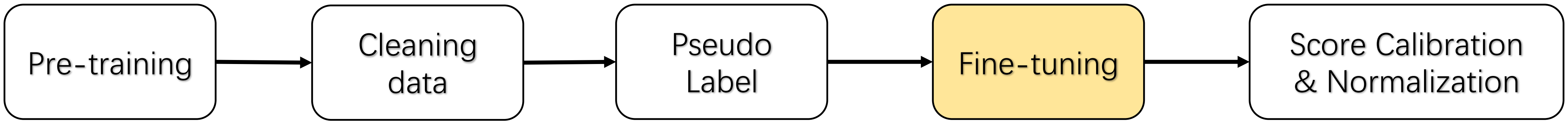
$$SSE = \sum_{i=1}^{k} \sum_{a=1}^{A} ||z_{k,a} - C_k||^2$$

`elbow' method to determine K

Among {1600,1800,2000} classes numbers, 1800 prevails.

Estimate cluster numbers and generate pseudo labels

4 Fine-tuning



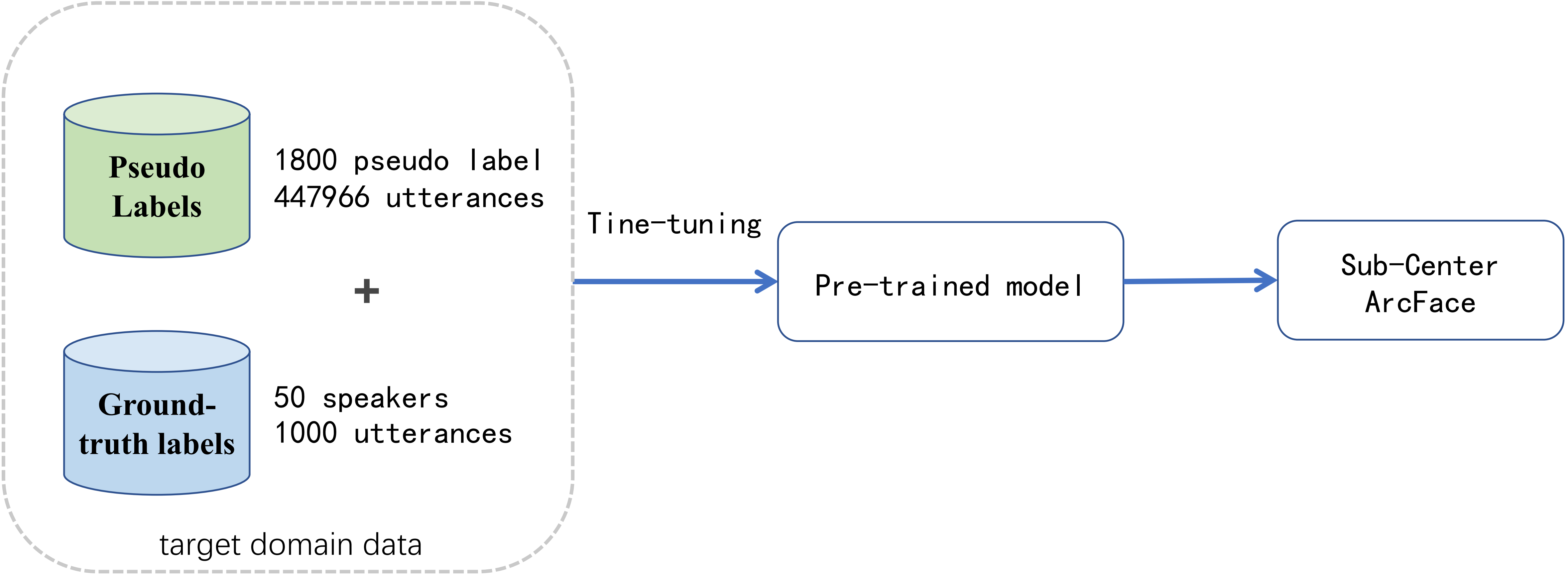
Pre-training

Cleaning data

Pseudo Label

Fine-tuning

Score Calibration & Normalization



Pseudo Labels

1800 pseudo label
447966 utterances

+

Ground-truth labels

50 speakers
1000 utterances

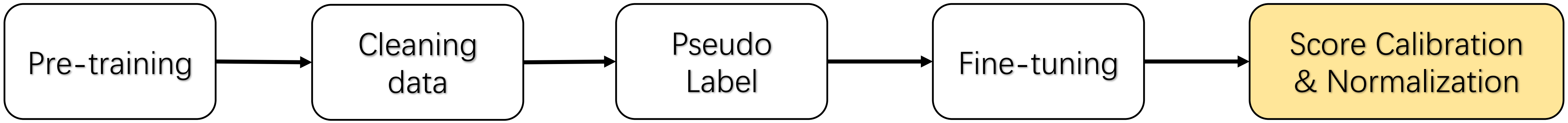
target domain data

Tine-tuning

Pre-trained model

Sub-Center ArcFace

5 Back-end



Pre-training

Cleaning data

Pseudo Label

Fine-tuning

Score Calibration & Normalization

Scoring :

Cosine similarity

Utterance-level AS-Norm:

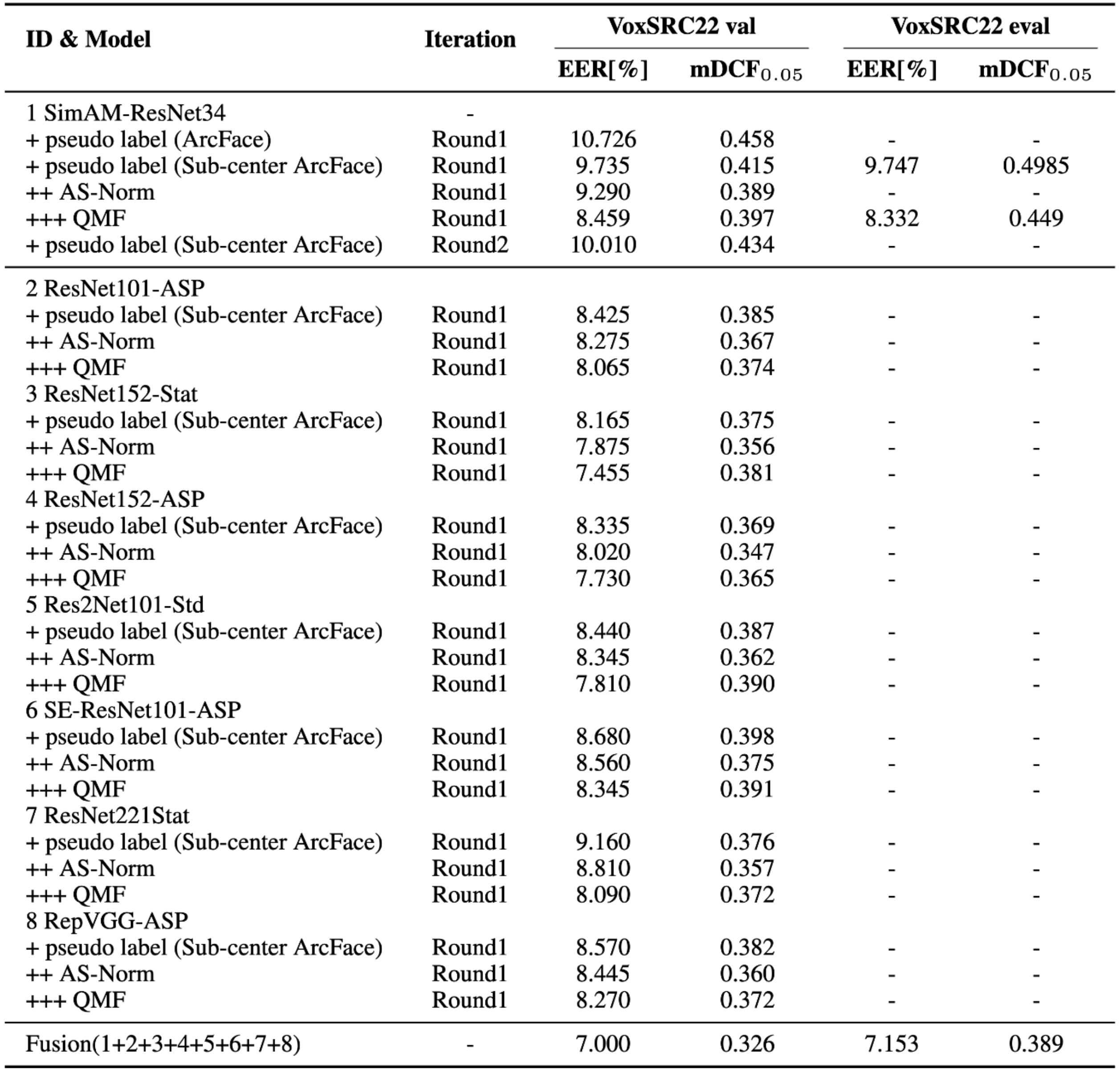
randomly select 20,000 utterances (duration over 4s) from unlabeled data as cohort set.

QMF:

speech duration: $log(d_u - d_{min})$

magnitude rate: $|log(\frac{||\mathbf{z}_e||}{||\mathbf{z}_t||})|$

Results



ID & Model	Iteration	VoxSRC22 val		VoxSRC22 eval	
		EER[%]	mDCF$_{0.05}$	EER[%]	mDCF$_{0.05}$
1 SimAM-ResNet34	-				
+ pseudo label (ArcFace)	Round1	10.726	0.458	-	-
+ pseudo label (Sub-center ArcFace)	Round1	9.735	0.415	9.747	0.4985
++ AS-Norm	Round1	9.290	0.389	-	-
+++ QMF	Round1	8.459	0.397	8.332	0.449
+ pseudo label (Sub-center ArcFace)	Round2	10.010	0.434	-	-
2 ResNet101-ASP					
+ pseudo label (Sub-center ArcFace)	Round1	8.425	0.385	-	-
++ AS-Norm	Round1	8.275	0.367	-	-
+++ QMF	Round1	8.065	0.374	-	-
3 ResNet152-Stat					
+ pseudo label (Sub-center ArcFace)	Round1	8.165	0.375	-	-
++ AS-Norm	Round1	7.875	0.356	-	-
+++ QMF	Round1	7.455	0.381	-	-
4 ResNet152-ASP					
+ pseudo label (Sub-center ArcFace)	Round1	8.335	0.369	-	-
++ AS-Norm	Round1	8.020	0.347	-	-
+++ QMF	Round1	7.730	0.365	-	-
5 Res2Net101-Std					
+ pseudo label (Sub-center ArcFace)	Round1	8.440	0.387	-	-
++ AS-Norm	Round1	8.345	0.362	-	-
+++ QMF	Round1	7.810	0.390	-	-
6 SE-ResNet101-ASP					
+ pseudo label (Sub-center ArcFace)	Round1	8.680	0.398	-	-
++ AS-Norm	Round1	8.560	0.375	-	-
+++ QMF	Round1	8.345	0.391	-	-
7 ResNet221Stat					
+ pseudo label (Sub-center ArcFace)	Round1	9.160	0.376	-	-
++ AS-Norm	Round1	8.810	0.357	-	-
+++ QMF	Round1	8.090	0.372	-	-
8 RepVGG-ASP					
+ pseudo label (Sub-center ArcFace)	Round1	8.570	0.382	-	-
++ AS-Norm	Round1	8.445	0.360	-	-
+++ QMF	Round1	8.270	0.372	-	-
Fusion(1+2+3+4+5+6+7+8)	-	7.000	0.326	7.153	0.389

Challenge summary:

- Data distribution of Validation set is same as evaluation set
- Multiple iteration is not work, only adopt one round of clustering and fine-tuning
- FT-domain is batter than FT-Mix
- AS-Norm has improvement
- QMF has great improvement
- Sub-center ArcFace is batter then ArcFace
- Purify pseudo label is not work (?)

FT-Mix (Vox2Dev together with CN-Celeb)
FT-domain (CN-Celeb)

The end

ming.li369@duke.edu
xiaoyi.qin@dukekunshan.edu.cn