

C-P MAP: A Novel Evaluation Toolkit for Speaker Verification

Lantian Li

CSLT @ THU

2022.3.14

ASV in deep learning era

- Backbones

ResNet

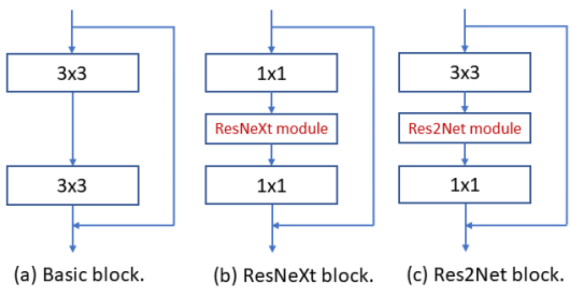


Fig. 1: *Three types of residual blocks.*

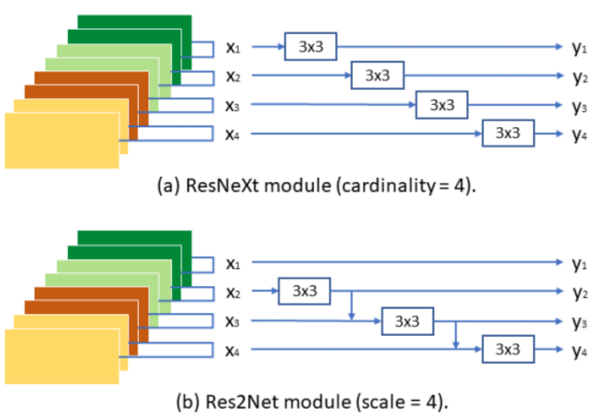
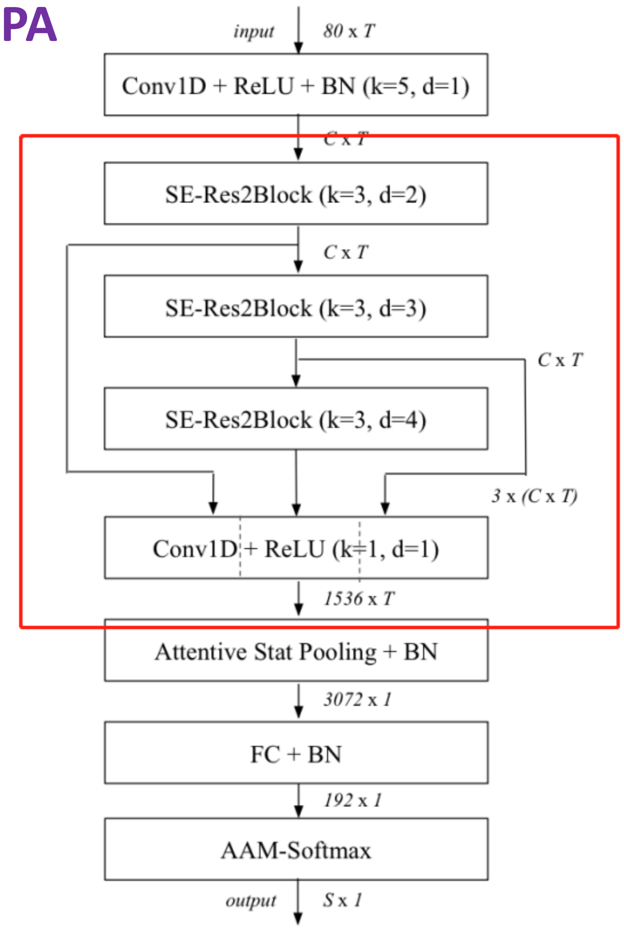


Fig. 2: *Detailed designs inside ResNeXt and Res2Net blocks..*

ECAPA



ASV in deep learning era

- Pooling strategies

TSP

$$\mathbf{m} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t$$

$$\mathbf{d} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \odot \mathbf{h}_t - \mathbf{m}\odot\mathbf{m}}$$

ASP

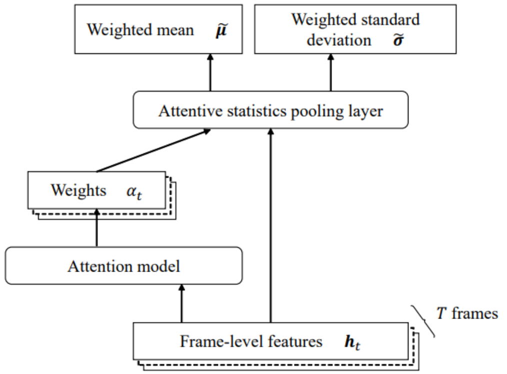
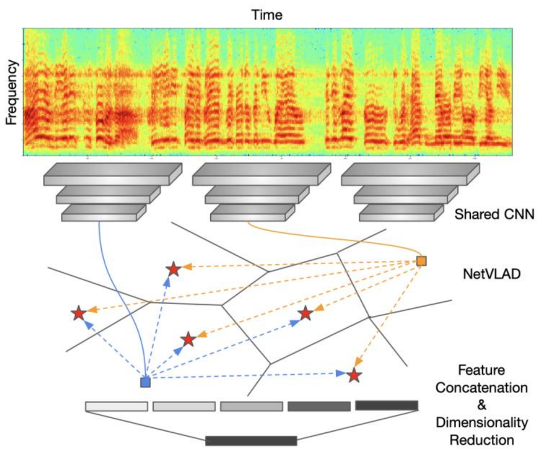
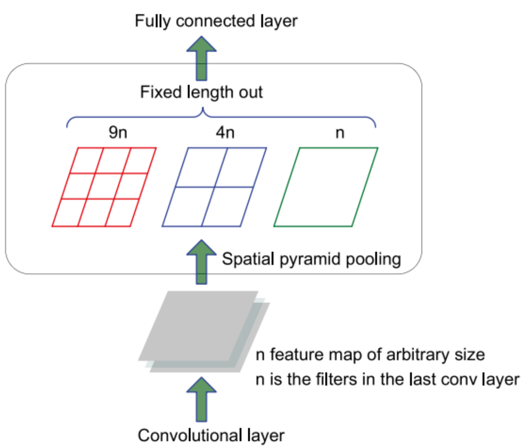


Figure 2: *Attentive statistics pooling*

VLAD

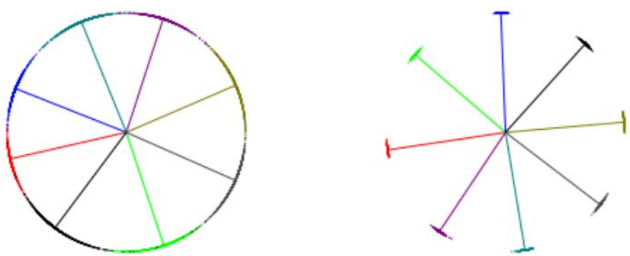


SPP



ASV in deep learning era

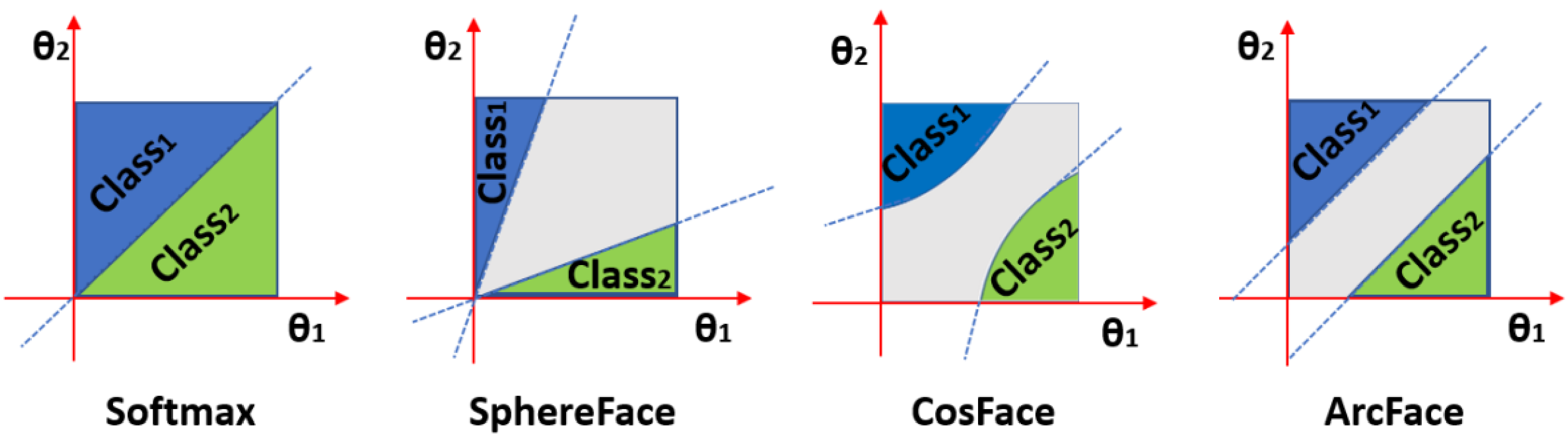
- Angular margin loss



(a) Softmax (b) ArcFace

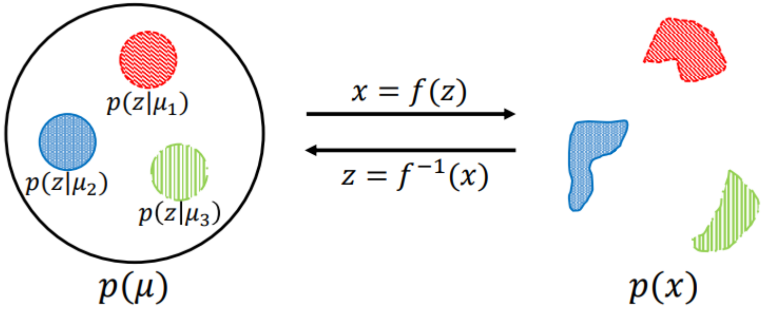
$$\psi(\theta_{y_i}) = \cos(m_1\theta_{y_i} + m_2) - m_3$$

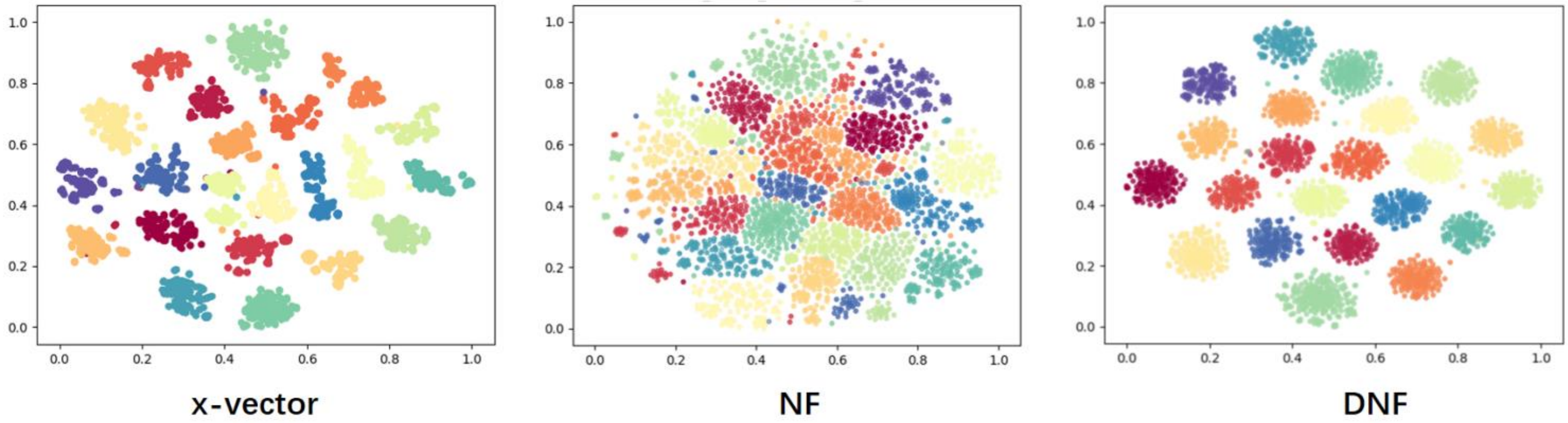
$$L_{LMS} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\psi(\theta_{y_i})}}{e^{s\cdot\psi(\theta_{y_i})} + \sum_{j=1,j\neq i}^{C} e^{s\cdot\cos\theta_j}}$$



ASV in deep learning era

- Score normalization





Impressive performance

- VoxSRC 2020

Track	Rank	Team Name	Organization	minDCF	EER
1	-	Baseline	Provided	0.477	7.68
	3	ntorgashov [15]	ID R&D Inc., New York, USA	0.203	3.82
	2	xx205 [16]	AISpeech Ltd, China	0.196	3.81
	1	JTBD [17]	IDLab, Ghent University, Belgium	0.177	3.73
2	-	Baseline	Provided	0.477	7.68
	3	DKU-DukeECE [18]	Duke Kunshan University, China & Duke University, USA	0.205	3.88
	2	xx205 [16]	AISpeech Ltd, China	0.194	3.80
	1	JTBD [17]	IDLab, Ghent University, Belgium	0.174	3.58
3	-	Baseline	Provided	0.877	19.07
	3	umair.khan [19]	TALP Research Center, UPC, Spain	0.751	14.71
	2	DKU-DukeECE [18]	Duke Kunshan University, China & Duke University, USA	0.598	12.42
	1	JTBD [17]	IDLab, Ghent University, Belgium	0.345	7.21

Impressive performance

- VoxSRC 2021

Track 1

#	User	Entries	Date of Last Entry	DCF ▲	EER ▲
1	snowstar	5	09/02/21	0.1034 (1)	1.8460 (1)
2	yugi	4	09/02/21	0.1175 (2)	2.8400 (3)
3	JTBD	5	09/02/21	0.1291 (3)	2.2710 (2)

Track 2

#	User	Entries	Date of Last Entry	DCF ▲	EER ▲
1	snowstar	5	09/02/21	0.1034 (1)	1.8460 (1)
2	yugi	4	09/02/21	0.1175 (2)	2.8400 (5)
3	JTBD	5	09/01/21	0.1313 (3)	2.0490 (2)

Benchmark vs. Deployment

ICS 03. 060
A11

JR

中 华 人 民 共 和 国 金 融 行 业 标 准

JR/T 0164—2018

移动金融基于声纹识别的安全应用
技术规范

Technical specifications for voiceprint recognition based
security application for mobile finance

6. 1 基本性能指标

基本性能指标应满足以下要求：
——错误接受率(FAR)≤0. 5%。
——错误拒绝率(FRR)≤3. 0%。

Deployment Performance

EER > 5.0%

- Benchmark-deployment Gap！

To interpret and settle this gap

- **Data theme**: hypothesizing that the performance gap is largely attributed to *acoustic mismatch*.
  - HI-MIA: Near-far filed mismatch
  - NIST SRE: Long-short mismatch, channel mismatch
  - VoxCeleb: Session mismatch
  - CN-Celeb: Genre mismatch
  - ……

Topology	Pooling	Loss	SITW	CN-Celeb.E
TDNN	TSP	Softmax	2.43	16.87
TDNN	TSP	AAM-Softmax	2.49	16.65
TDNN	SAP	Softmax	2.41	17.11
TDNN	SAP	AAM-Softmax	2.57	16.96
ResNet-34	TSP	Softmax	2.41	16.74
ResNet-34	TSP	AAM-Softmax	1.96	16.51
ResNet-34	SAP	Softmax	2.16	17.33
ResNet-34	SAP	AAM-Softmax	2.30	16.52

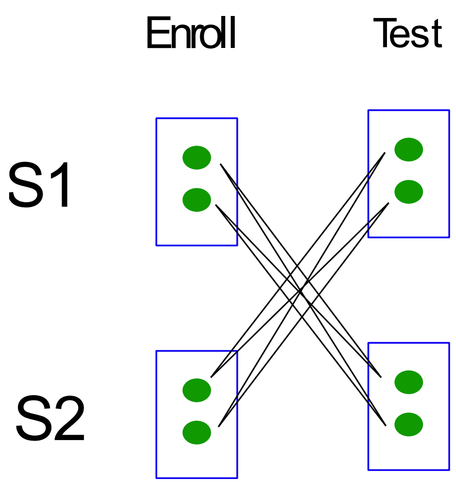
To interpret and settle this gap

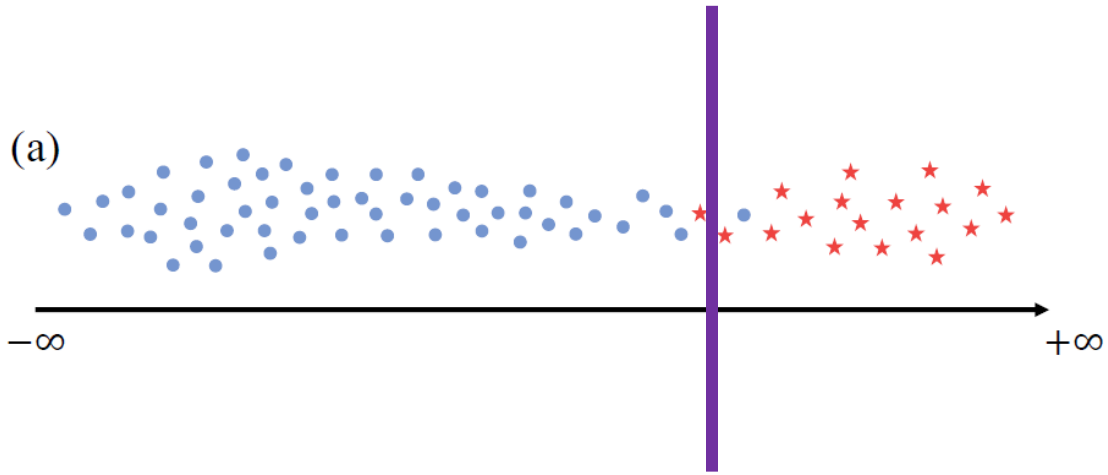
- **Trials theme**
  - Each trial is an individual test case.

  - We argue that there is the *bias* on evaluation trials, leading to the benchmark-deployment gap.

Cross-pairing trials design

- For example, cross-pairing design produces a larger proportion of *easy* trials, leading to over optimistic performance estimation.
- Target trials: NK(K-1) vs. Negative trials: N(N-1)K^2





Cross-pairing vs. Real-life

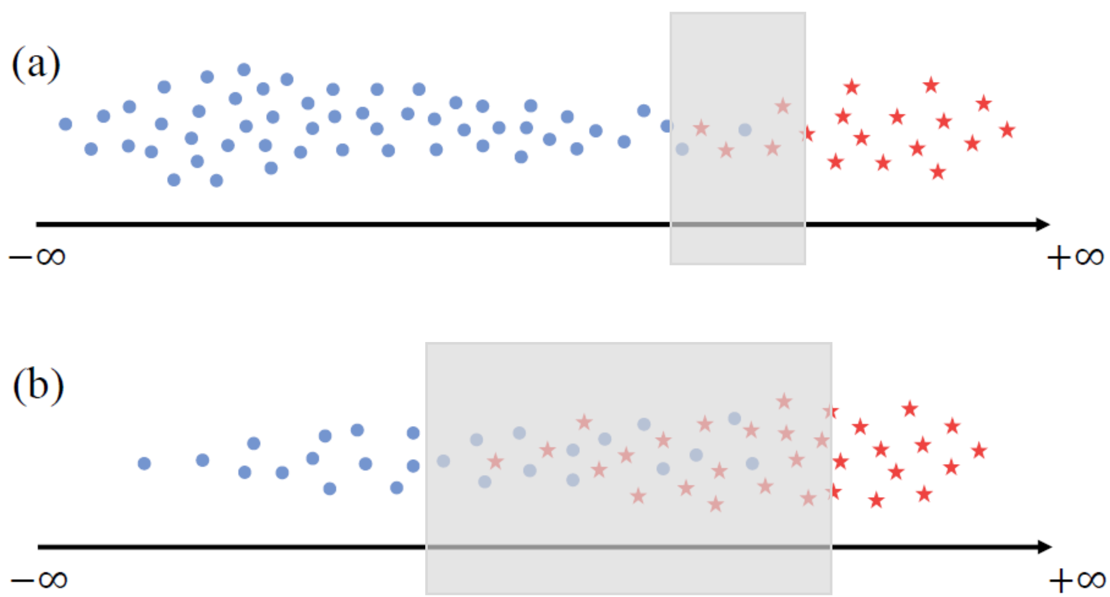
- Cross-pairing trials
  - There is a large proportion of *easy* trials, particularly the cases for negative trials.
  - More negative trials than positive trials.

- Real-life trials
  - The negative trials more challenging as the imposters often with the same acoustic condition, such as gender, accent, language.
  - More positive trials than negative trials.

Trial bias issue

(a) shows the scores of trials created by cross-pairing.

(b) shows the scores of trials encountered in real-life.



- The distribution difference reflects the bias on trials.

Concept of Trial config

- Given a set of enrollment/test utterances, **a trial config** is defined as *a subset of trials selected to test against an ASV system.*

- The full cross-pairing is the largest trial config and involves all the possible trials.

- For an ASV system, performance with different trial configs are different, reflecting real performance under different deployment conditions.

Config-performance map

- By collecting **all possible trial configs** and computing the corresponding performance, we can evaluate the ASV system in a more thorough way.

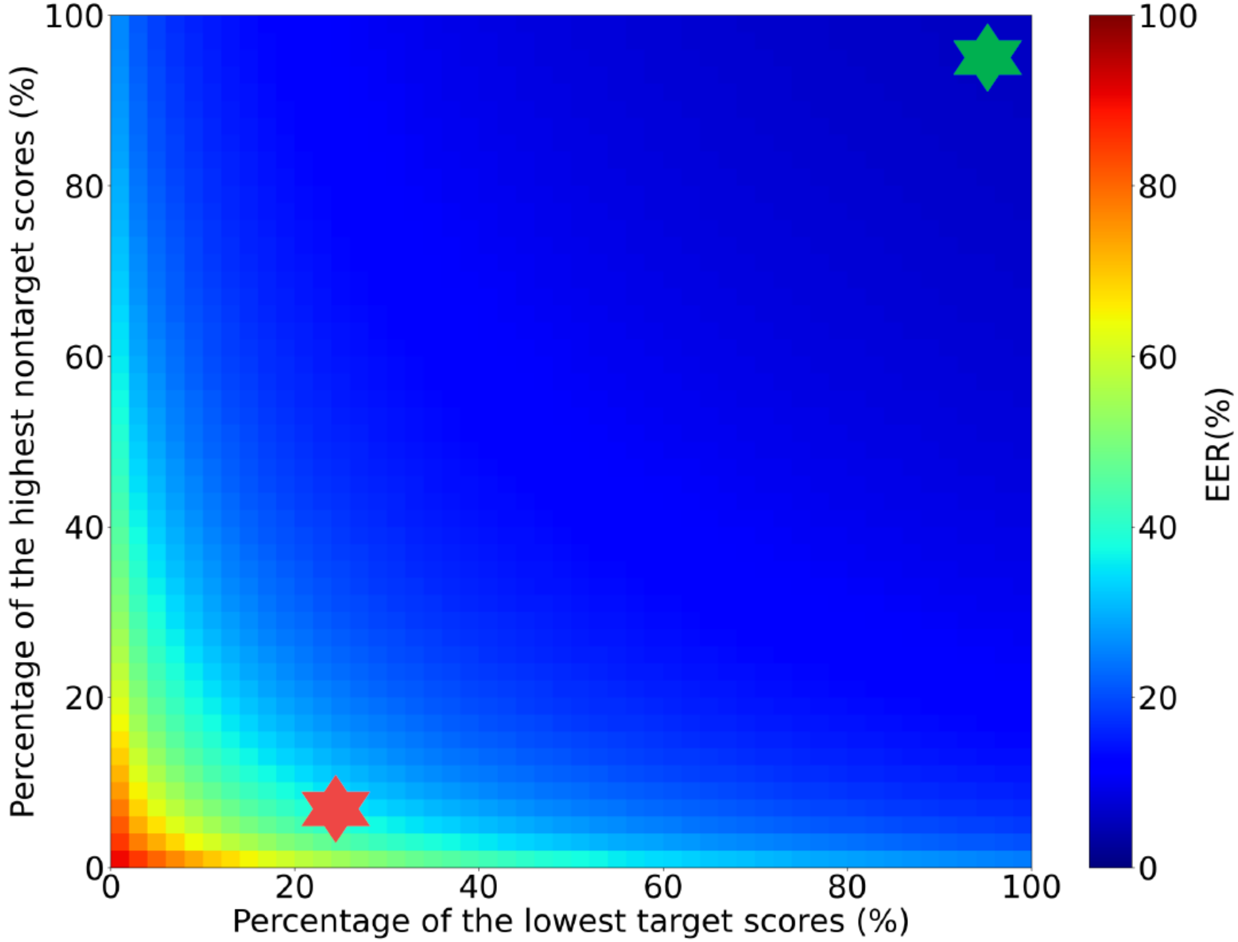
- The process of C-P map
  - x-axis corresponds to subsets of positive trials.
  - y-axis corresponds to subsets of negative trials.
  - each location (x; y) on the map corresponds to a particular trial config.
  - The color at (x; y) represents the performance.

Take an example

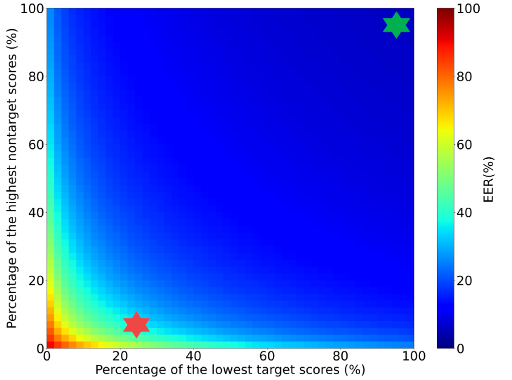
- Score-ordered trial configs
  - Target trials sets (x-axis): trials with *higher scores from left to right.* **[hard to easy]**
  - Non-target trials sets (y-axis): trials with *lower scores from bottom to up.* **[hard to easy]**

- The color in the map represents the EER values corresponding to each trial config.

C-P map of the i-vector system



Observations

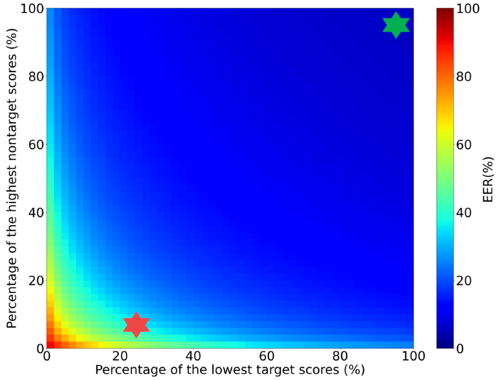


- The large proportion of high-performance area reveals that there are larger amount of easy trials.

- Two trial configs (red star and green star) represents the real-life deployment and the cross-pairing benchmark.

- It is clear that the two trial configs lead to quite different EER results, which is precisely the benchmark-deployment gap.

The value of C-P map



- If the order of the trial configs are fixed, the C-P map is more useful.

- System analysis and comparison
  - Create ordered trial configs by fusing several basic systems.
  - With these trial configs, we can plot C-P maps for an ASV systems to obtain detailed analysis.
  - Moreover, we can plot the relative change between two systems for system comparison.

Basic systems

- Data
  - Training set: *VoxCeleb2.dev*
  - Evaluation set: *VoxCeleb1-O* and *VoxCeleb1-E*

- Basic system
  - i-vector and x-vector

- More powerful systems
  - ResNet34, Attentive pooling, AM-Softmax

System performance

Table 1: EER(%) and minDCF results with the modern ASV systems on VoxCeleb1 evaluation trials.

System	Front-End	Back-End	VoxCeleb1-O		VoxCeleb1-E	
			EER(%)	minDCF	EER(%)	minDCF
1	GMM i-vector	PLDA	5.819	0.5189	5.872	0.5038
2	TDNN + TSP + Softmax	PLDA	4.558	0.4882	4.290	0.4343
3	TDNN + TSP + AM-Softmax	Cosine	3.430	0.3370	3.389	0.3619
4	ResNet34 + TSP + AM-Softmax	Cosine	1.633	0.1770	1.688	0.1900
5	ResNet34 + TSP + AAM-Softmax	Cosine	1.803	0.1961	1.747	0.1946
6	ResNet34 + ASP + AM-Softmax	Cosine	1.521	0.1642	1.504	0.1669

- Sys 1 and Sys 2 are used to produce trial configs.

C-P maps with EER metric

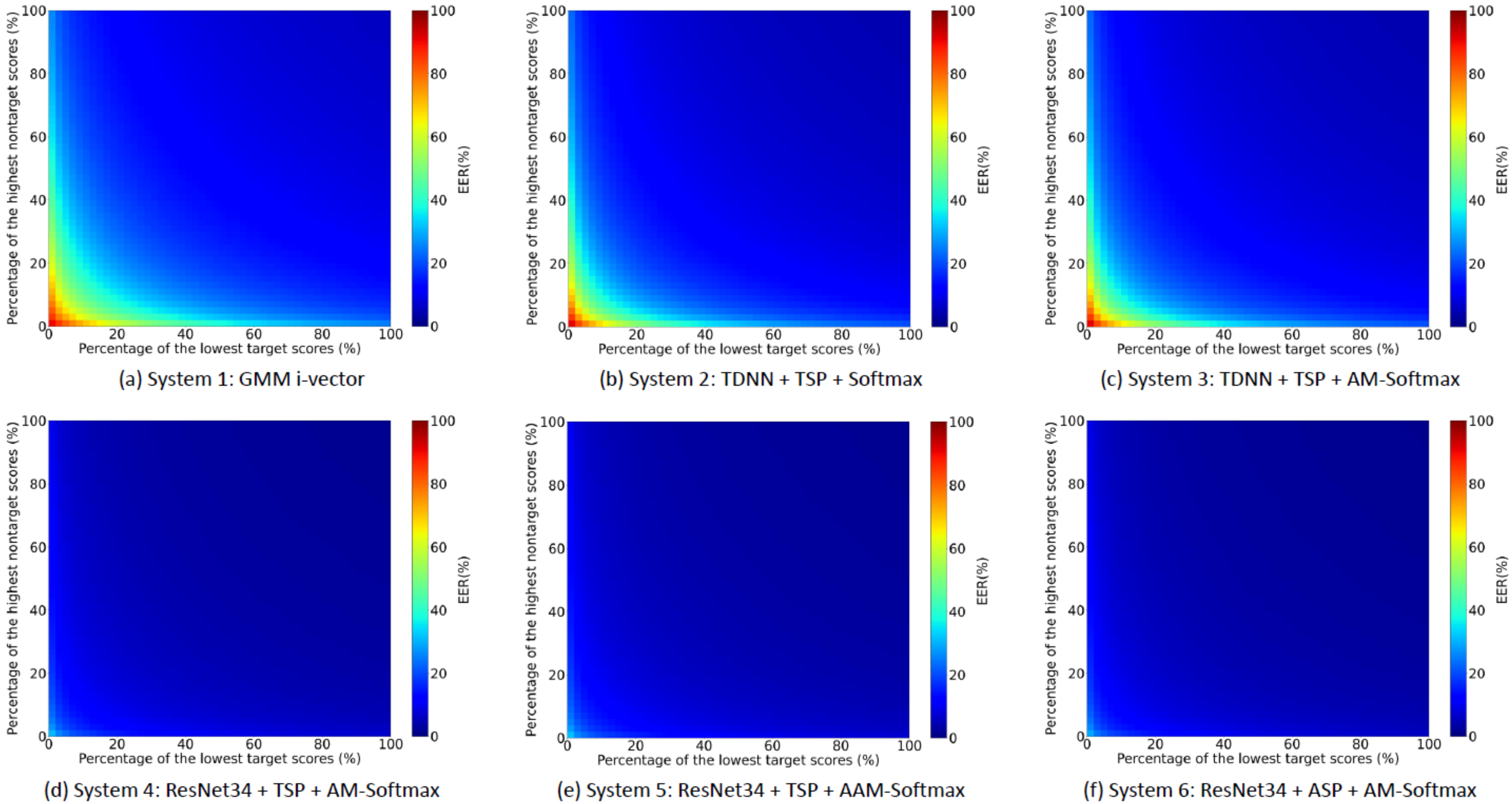
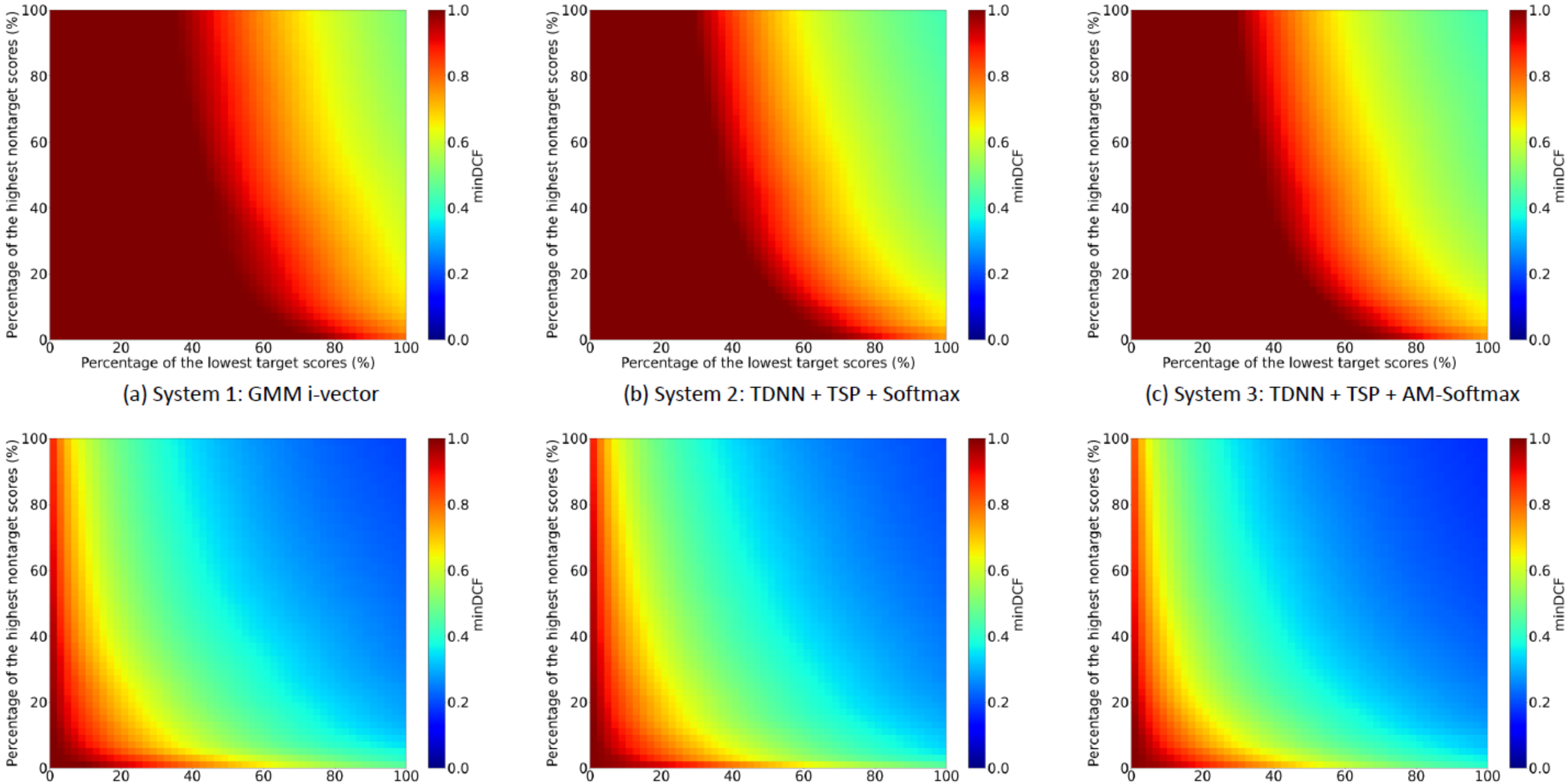


Figure 4: The C-P maps of 6 systems tested on VoxCeleb-E trials with *EER* metric.

C-P maps with minDCF metric



(a) System 1: GMM i-vector

(b) System 2: TDNN + TSP + Softmax

(c) System 3: TDNN + TSP + AM-Softmax

(d) System 4: ResNet34 + TSP + AM-Softmax

(e) System 5: ResNet34 + TSP + AAM-Softmax

(f) System 6: ResNet34 + ASP + AM-Softmax

Figure 5: The C-P maps of 6 systems tested on VoxCeleb-E trials with *minDCF* metric.

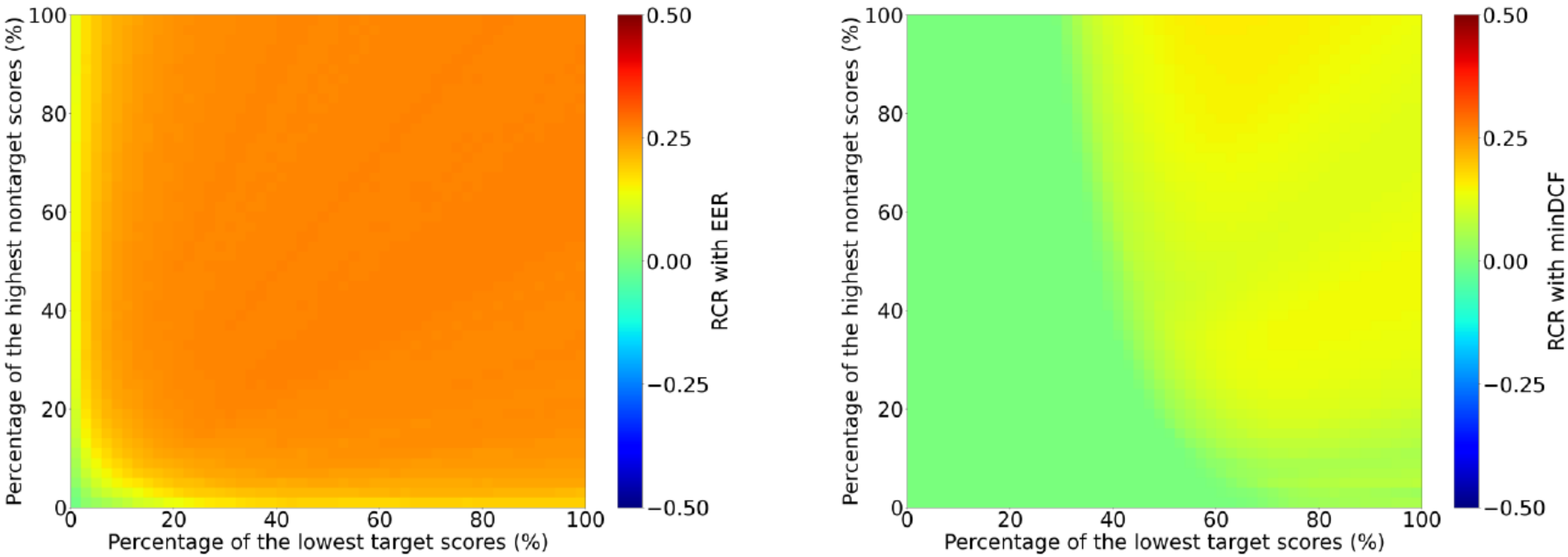
Delta C-P map

- The relative change ratio (RCR) at location (x,y) on two C-P maps.

$$\mathrm{RCR}(x,y) = \frac{\mathrm{CP}_{ref}(x,y) - \mathrm{CP}_{test}(x,y)}{\mathrm{CP}_{ref}(x,y)},$$

- If RCR > 0, it means the test system wins. If RCR < 0, it means the test system loses. If RCR = 0, they are tied.
- Win: Tie: Lose

i-vector vs. x-vector



Win : Tie : Lose = 99.96% : 0.00% : 0.04%

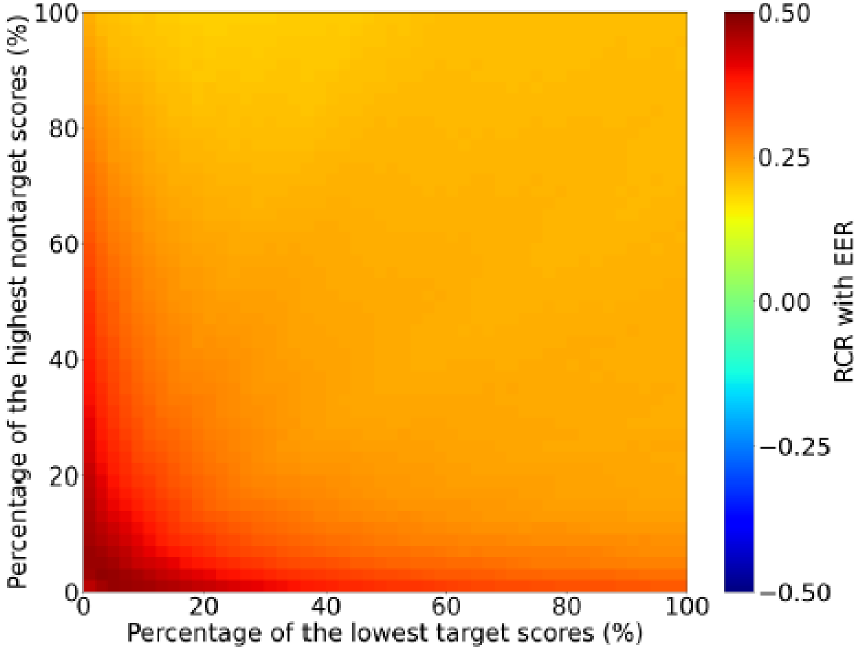
(a) x-vector against i-vector (*EER*)

Win : Tie : Lose = 60.68% : 29.96% : 9.36%

(b) x-vector against i-vector (*minDCF*)

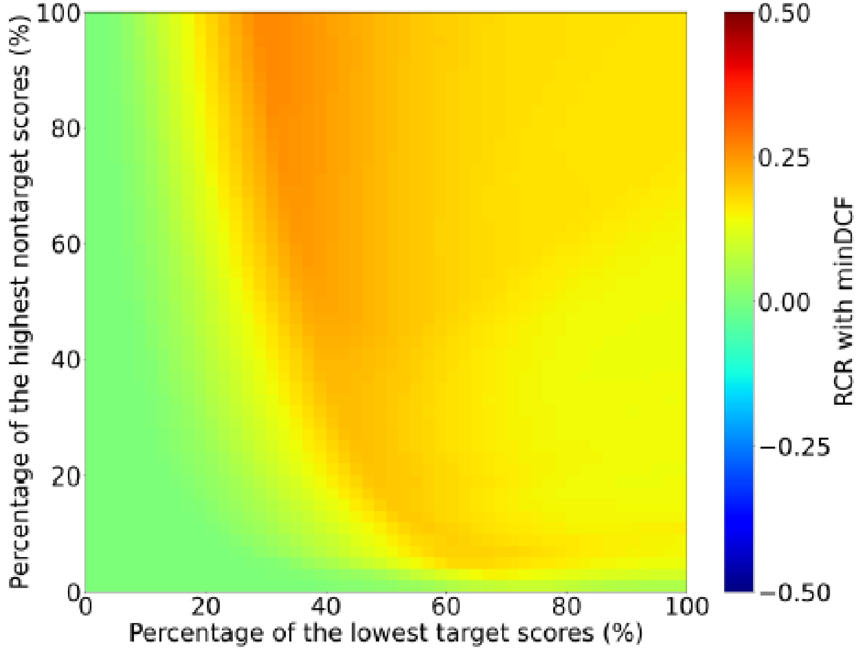
- The discriminative model is superior to the probabilistic model.

Softmax vs. AM-Softmax



Win : Tie : Lose = 100.00% : 0.00% : 0.00%

(a) AM-Softmax against Softmax (*EER*)

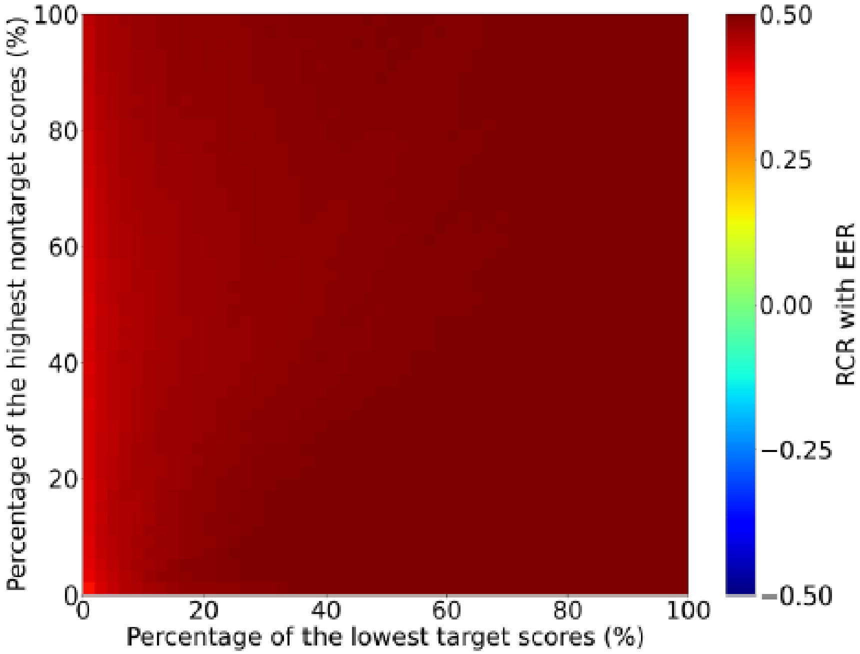


Win : Tie : Lose = 98.68% : 1.32% : 0.00%

(b) AM-Softmax against Softmax (*minDCF*)

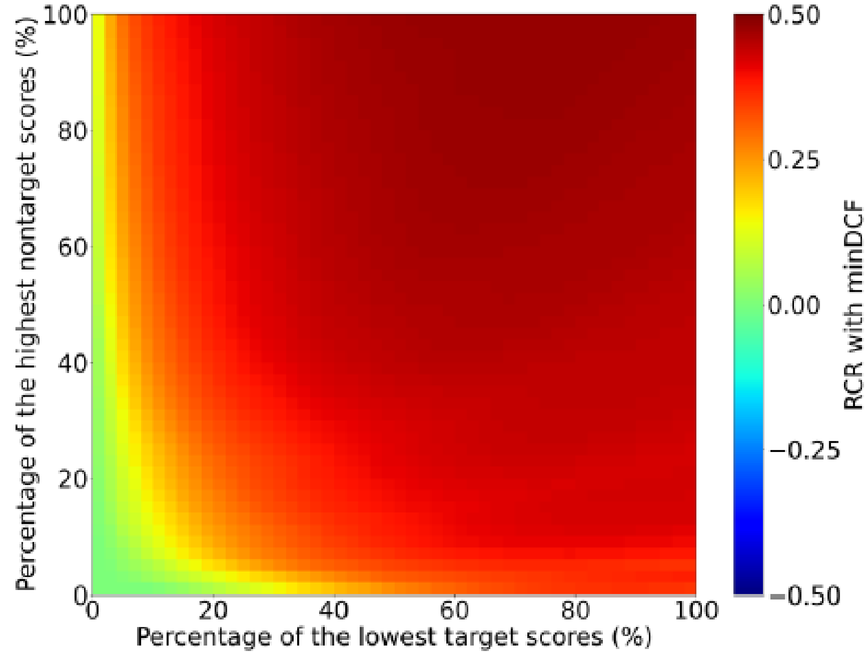
- The margin-based AM-Softmax overwhelmingly outperforms the standard Softmax.

TDNN vs. ResNet34



Win : Tie : Lose = 100.00% : 0.00% : 0.00%

(a) ResNet34 against TDNN (*EER*)

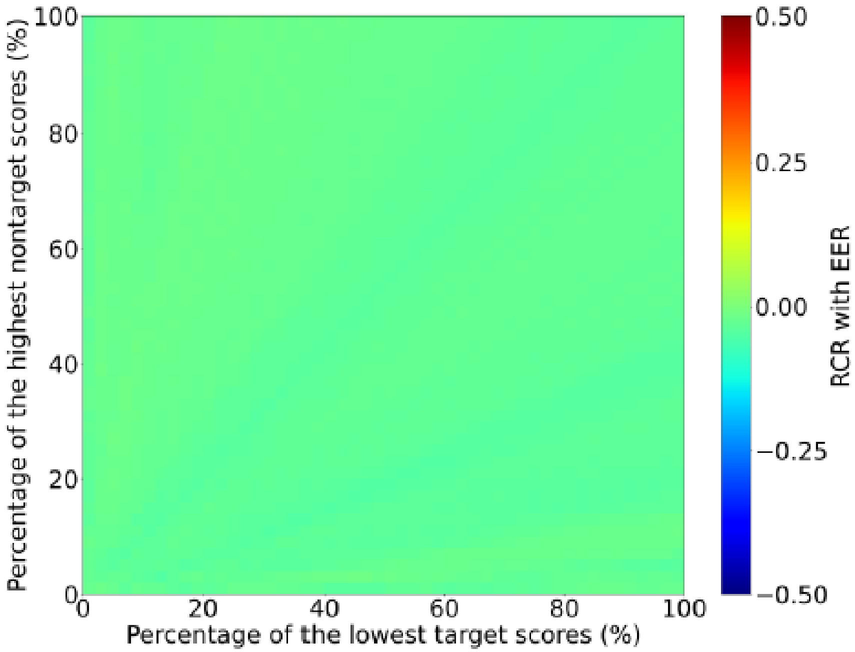


Win : Tie : Lose = 100.00% : 0.00% : 0.00%

(b) ResNet34 against TDNN (*minDCF*)

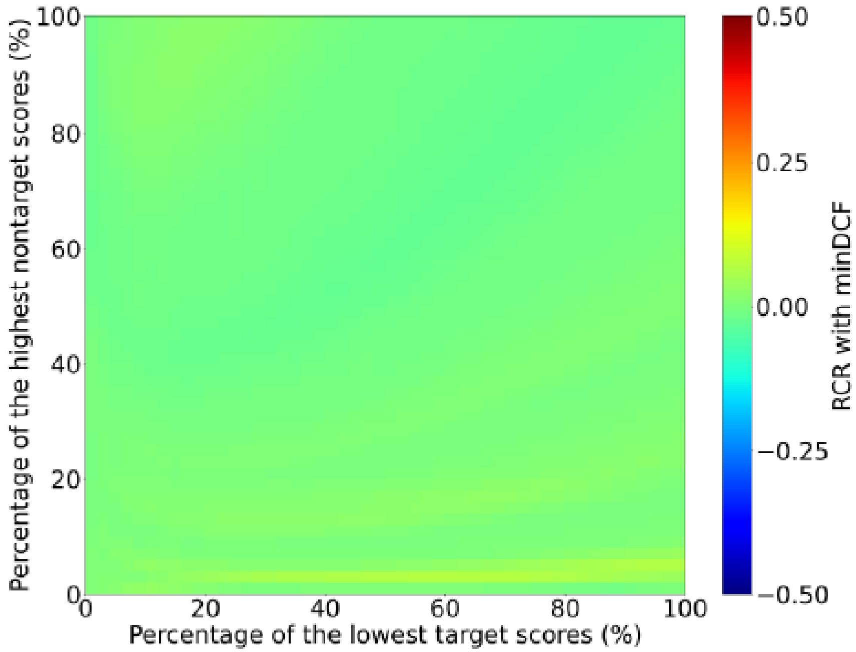
- It demonstrates the great success of ResNet34 in speaker recognition.

AAM-Softmax against AM-Softmax



Win : Tie : Lose = 0.00% : 0.00% : 100.00%

(a) AAM-Softmax against AM-Softmax (*EER*)

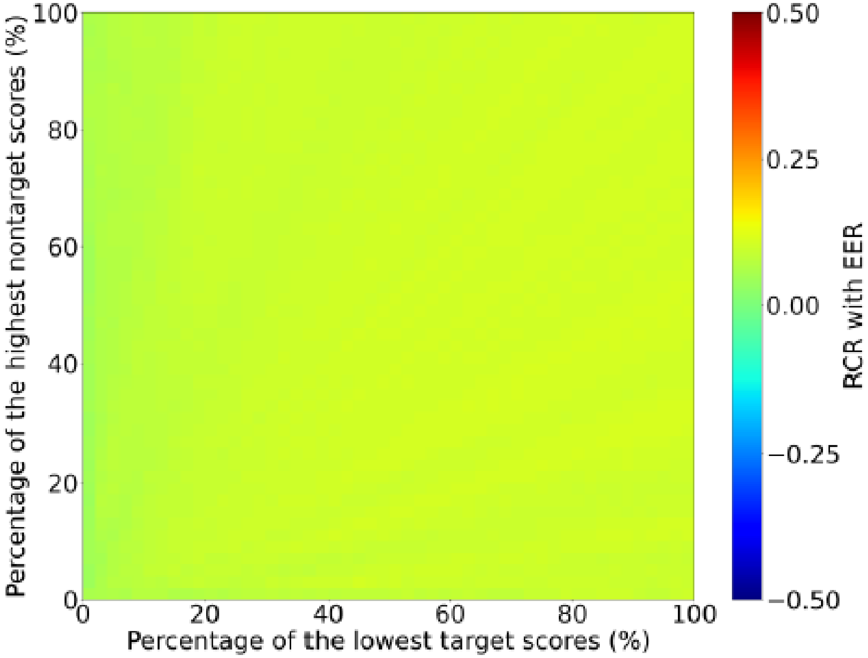


Win : Tie : Lose = 35.64% : 0.04% : 64.32%

(b) AAM-Softmax against AM-Softmax (*minDCF*)

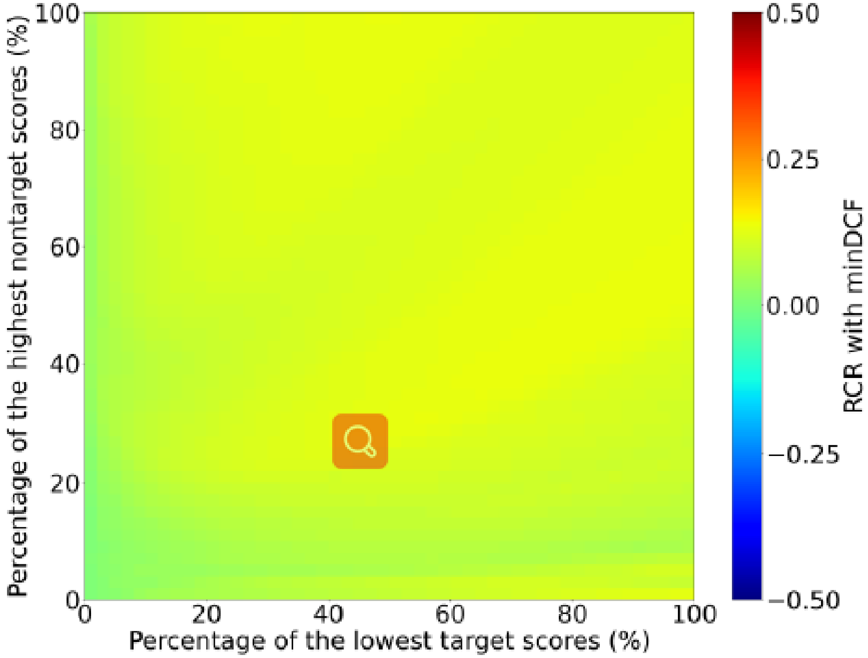
- The performance gap is quite marginal.

TSP vs. ASP



Win : Tie : Lose = 100.00% : 0.00% : 0.00%

(a) ASP against TSP (*EER*)



Win : Tie : Lose = 99.88% : 0.12% : 0.00%

(b) ASP against TSP (*minDCF*)

- ASP outperforms TSP on the whole.

Roadmap

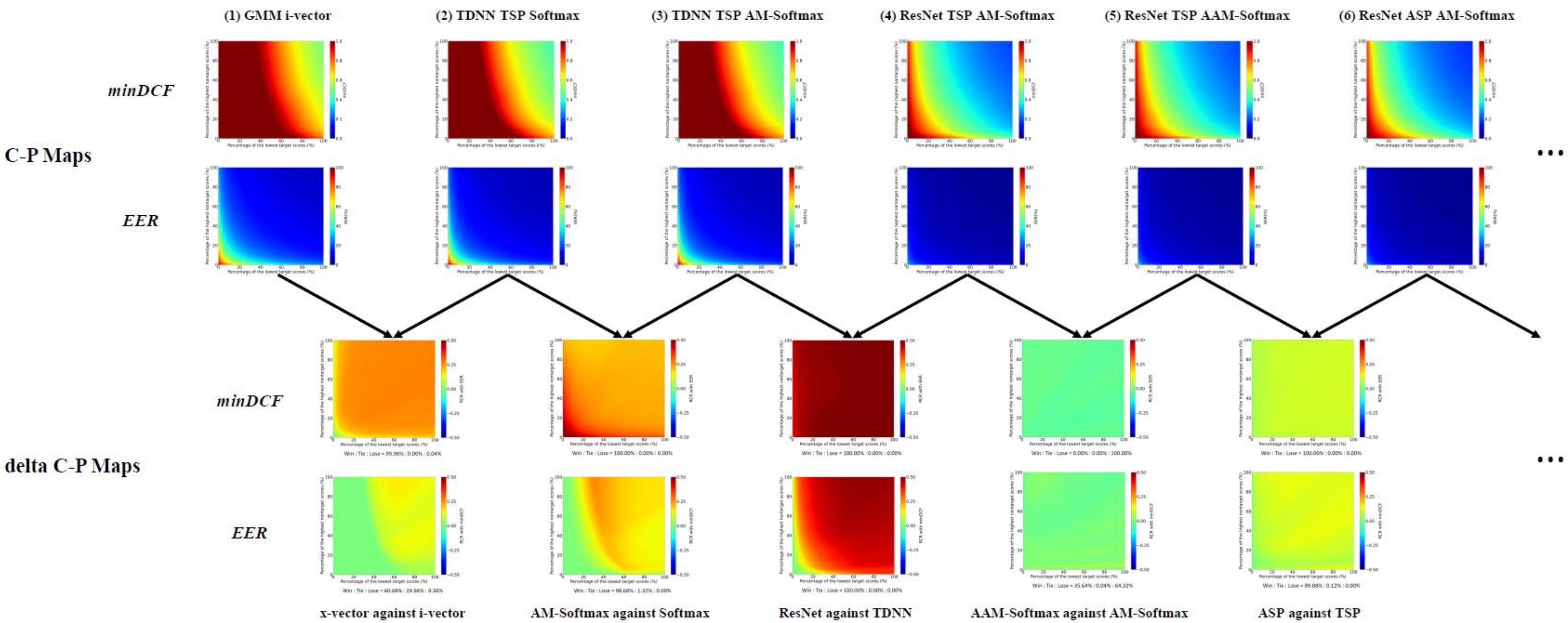


Figure 11: The roadmap of speaker recognition techniques measured by C-P map and delta C-P map.

Conclusions

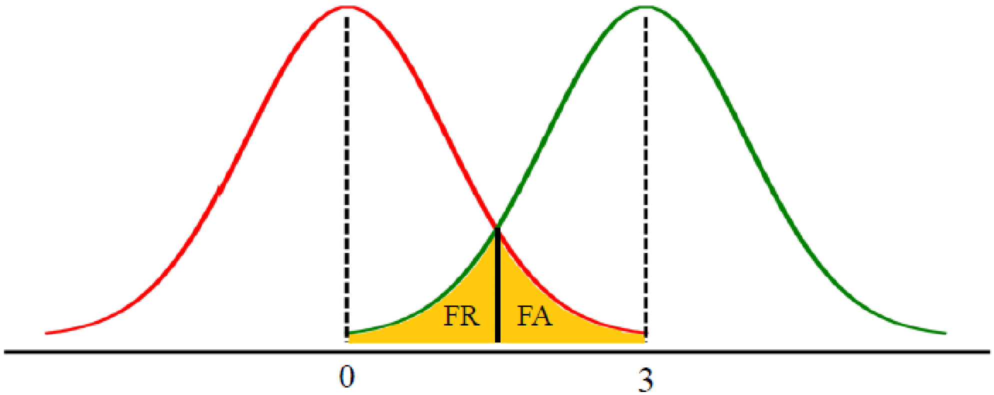
- This paper is inspired by the benchmark-deployment discrepancy.
- We hypothesize that this problem is attributed to the potential trial bias issue.
- To verify our hypothesis, we define the concept of trial config and its derived C-P map.
- We show that this C-P map is a novel evaluation tool for ASV system analysis and comparison.

Let us discuss one thing

- Are the performance measurements shown at different locations on the C-P map comparable?

- YES !

$$\int_{-\infty}^{\theta} p(c)\mathrm{d}c = \int_{\theta}^{+\infty} q(c)\mathrm{d}c$$



- The evaluation measurement (e.g., EER) are determined by distributions of scores of trials rather than trials themselves.