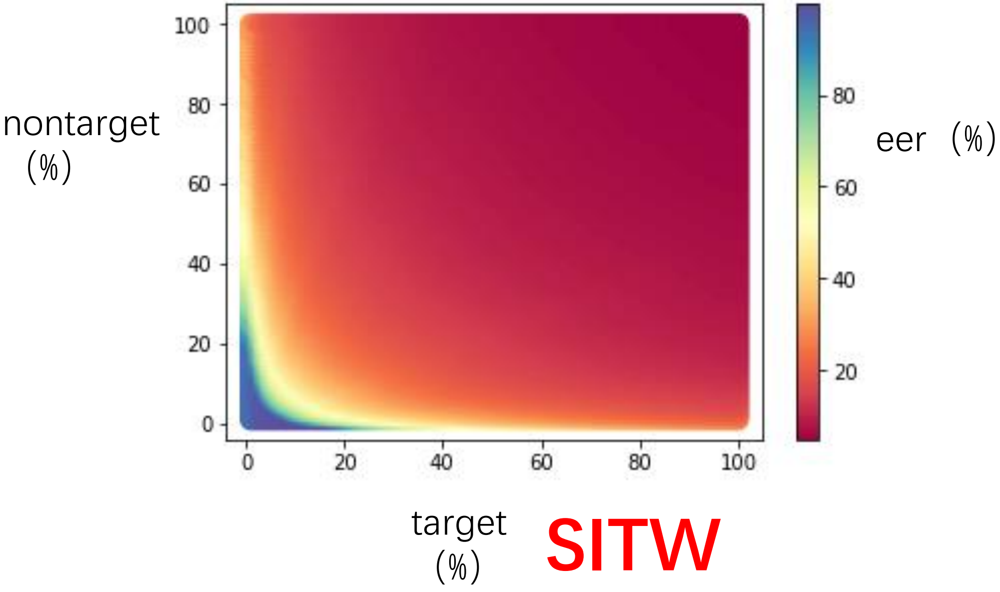
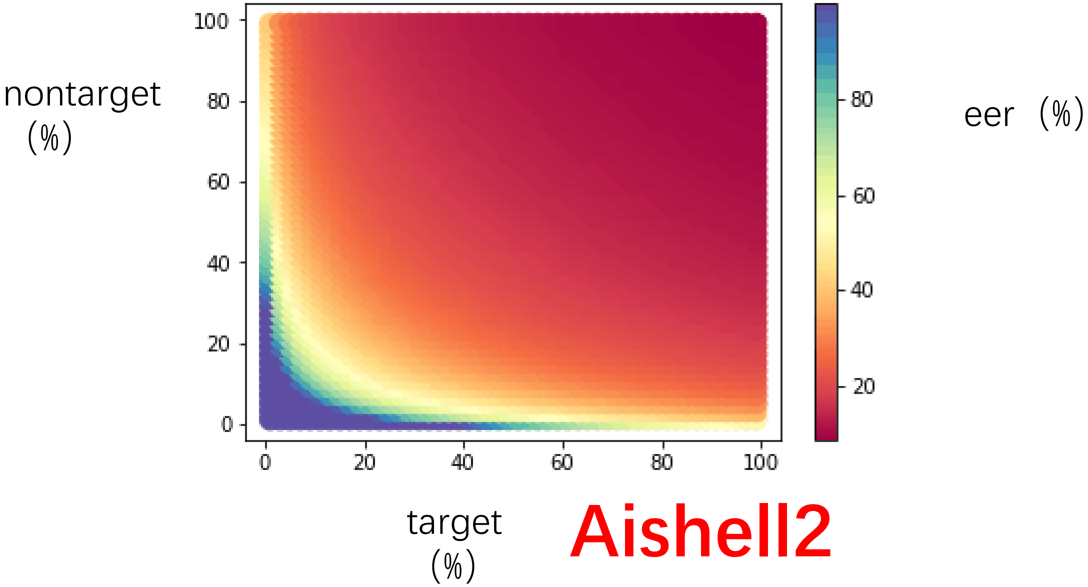
Secret of 'hard trials'

EER——在xvector模型中，取nontarget与target不同比例的trials进行plda打分

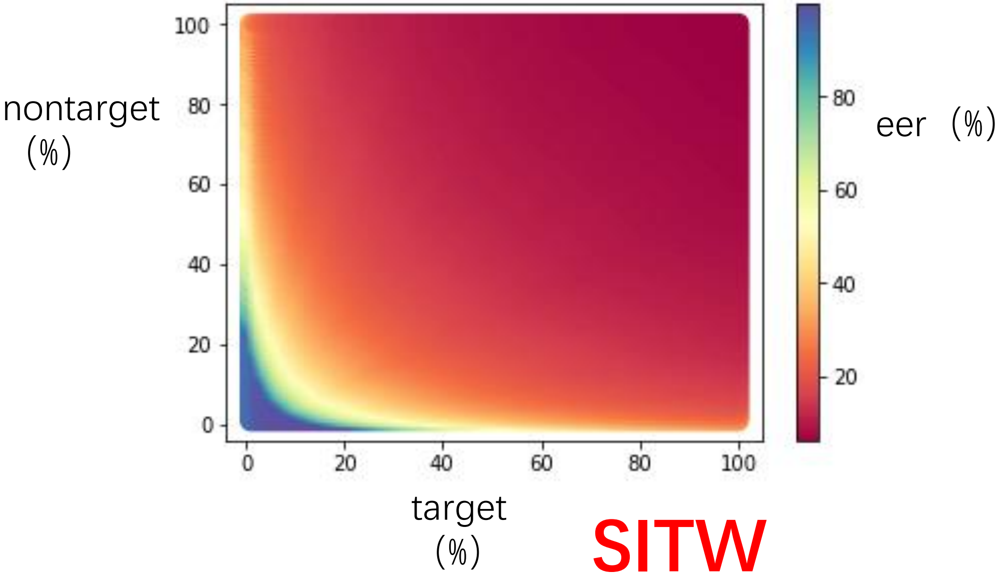


SITW

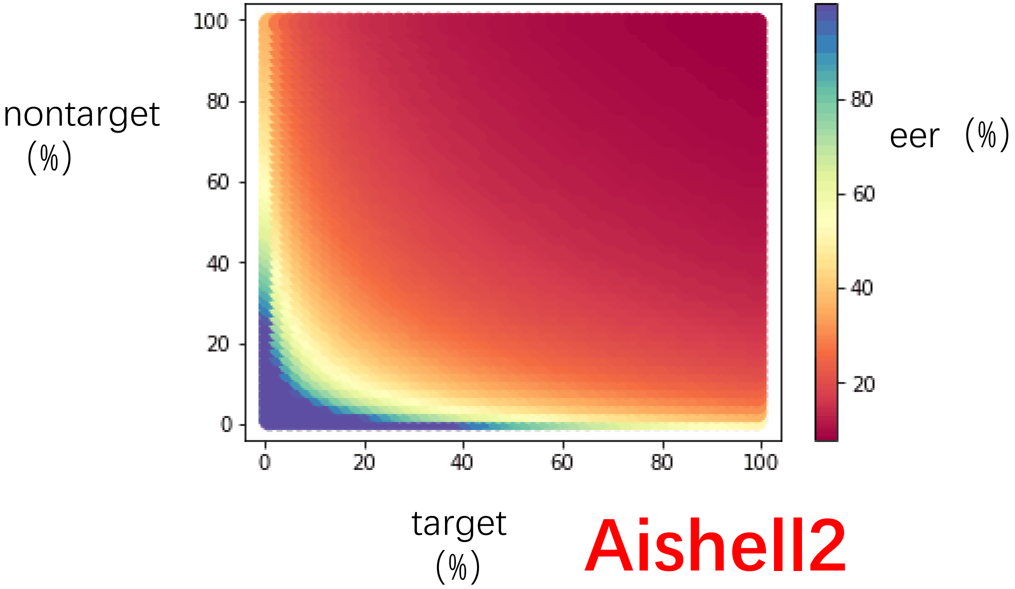


Aishell2

EER——在ivector模型中，取nontarget与target不同比例的trials进行plda打分



SITW



Aishell2

Current Situation

- It seems that the current deep speaker models (x-vector) have achieved the state-of-the-art performance on several benchmark datasets. For example, the EER performance in AISHELL-1(iOS) can achieve 0.73% with carefully tuned.

Really?

- In this work, we will make some investigations on the secret of 'hard trials'. Do hard trials really exisit ? If so, what do they sound like? Can we human identify them?

Preparation Work

- **Data preparation**
  - Training set: VoxCeleb2 dev, which contains 5994 speakers.

- **Speaker models**
  - Here we prepare two SOTA speaker embeddings, including i-vector and x-vector.
  - For each speaker embedding, we select 4 model configurations.

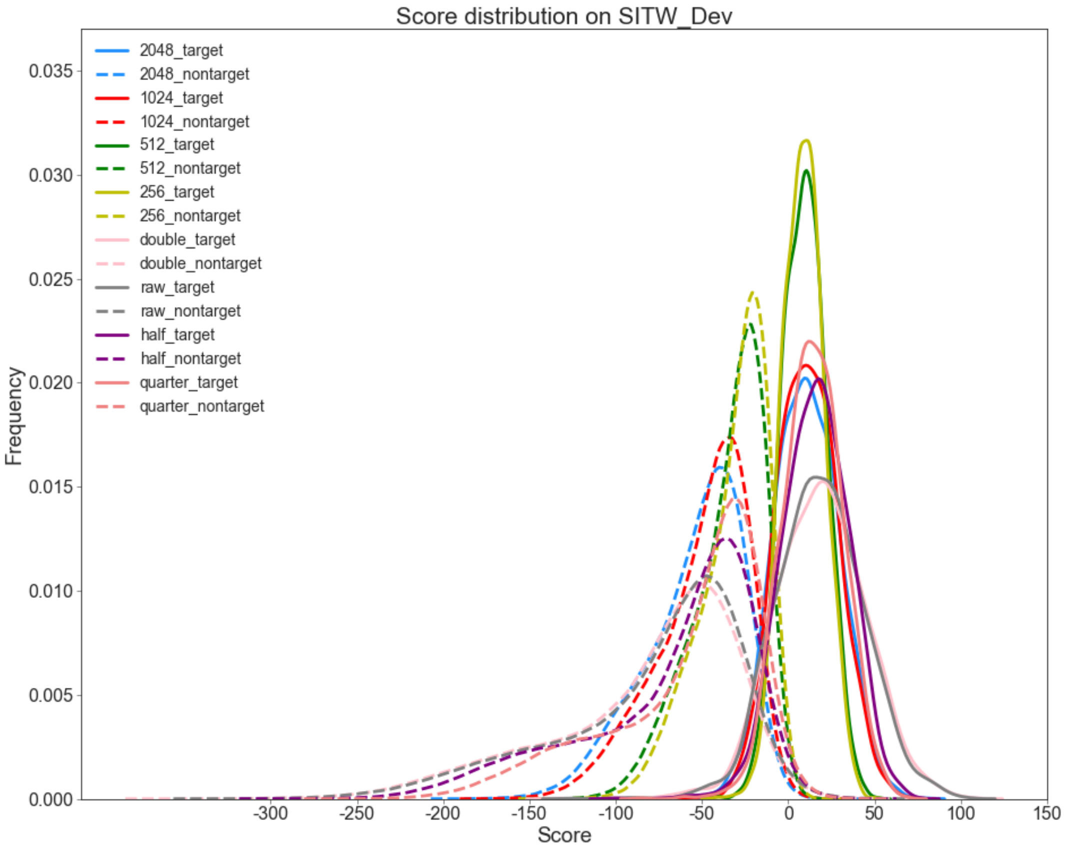
- **Evaluation trials**
    - There are 5 evaluation trials in total.
    - SITW.Dev.Core, SITW.Eval.Core, VoxSRC_O_cl, VoxSRC_E_cl, VoxSRC_H_cl

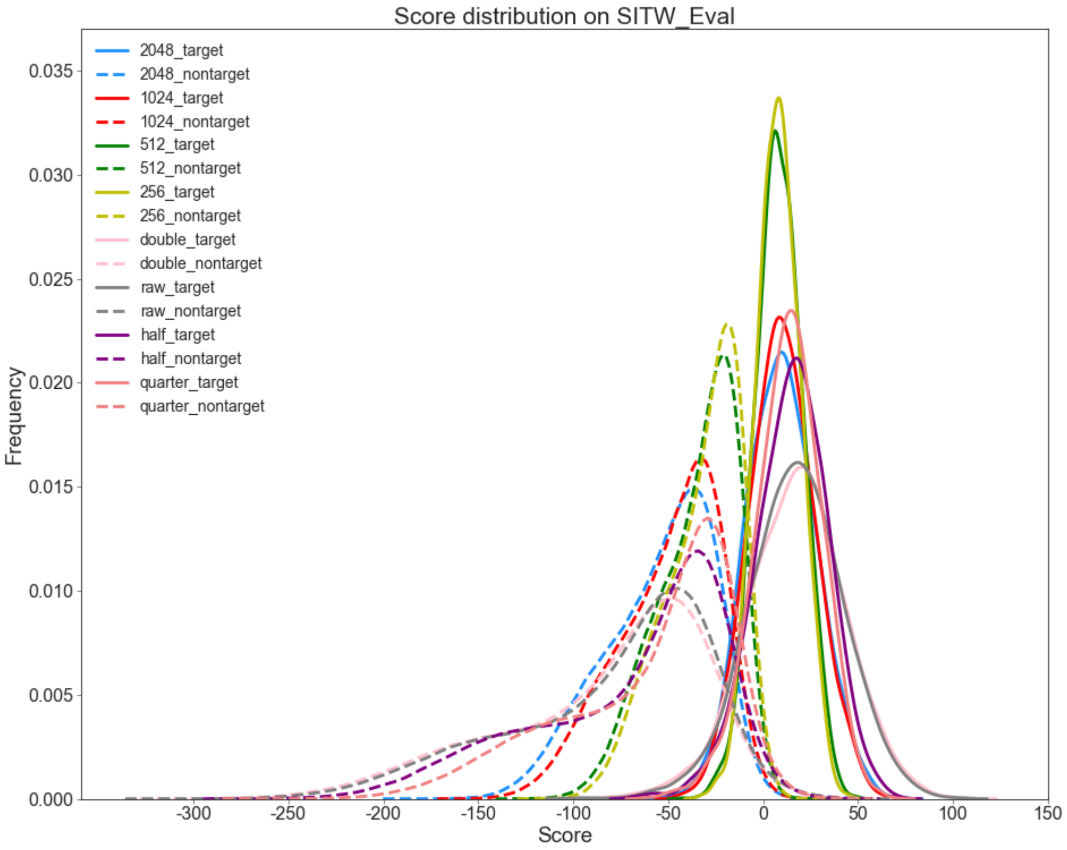
Basic results

PLDA EER（%）		VoxSRC			SITW	
	model_conf	trials	trials_E	trials_H	Dev Core	Eval Core
i-vector	2048_400	5.834（-16.79029）	5.899（-17.43673）	9.437（-13.93998）	5.853（-17.1198）	6.725（-16.87873）
	1024_400	6.095（-15.35659）	6.298（-16.21754）	10（-12.92543）	5.968（-14.86925）	7.19（-15.15215）
	512_200	7.095（-9.727581）	7.315（-10.35523）	11.4（-7.773234）	7.047（-8.735895）	7.736（-8.372811）
	256_200	7.834（-8.984595）	7.958（-9.479876）	12.33（-7.033204）	7.547（-7.613235）	8.748（-7.844908）
x-vector	double	5.116/5.111(-10.71275)	4.858（-12.3257）	7.633（-8.646959）	6.546（-18.95111）	7.682（-19.71514）
	raw	5.356/5.350（-10.68804）	5.086（-12.02205）	7.97（-8.426387）	6.662（-18.05646）	7.354（-17.99542）
	half	5.111(-10.20641)	4.93（-11.26423）	8.051（-7.748494）	6.585（-12.76487）	7.709（-13.58204）
	quarter	5.967/5.962（-8.232863）	5.797（-9.319304）	9.239（-6.357263）	7.008（-10.73184）	8.475（-11.62748）

Qualitative analysis

- 对于各个trials，绘制其在8个不同模型上的score分布图，来观察各个模型之间分数交叠的情况。





下表为在不同trials中，两个模型根据各自阈值选出重叠的hard trials，占两个模型各自hard trials之和的比重：

$$\frac{\text{overlap(hard trials)}}{\text{model\_A(hard trials)+ model\_B(hard trials)}}$$

Trials_E	2048	1024	512	256	double	raw	half	quarter
2048	50.000%							
1024	33.084%	50.000%						
512	31.035%	31.621%	50.000%					
256	29.725%	30.603%	33.689%	50.000%				
double	19.391%	19.149%	18.434%	17.976%	50.000%			
raw	19.844%	19.463%	18.721%	18.493%	28.970%	50.000%		
half	20.916%	20.534%	19.949%	19.432%	27.387%	28.044%	50.000%	
quarter	21.159%	20.990%	20.876%	20.629%	26.283%	27.053%	28.361%	50.000%

Trials_O	2048	1024	512	256	double	raw	half	quarter
2048	50.000%							
1024	33.378%	50.000%						
512	31.818%	32.373%	50.000%					
256	30.687%	31.348%	33.820%	50.000%				
double	19.140%	18.359%	18.463%	17.577%	50.000%			
raw	19.468%	18.997%	18.471%	18.105%	27.146%	50.000%		
half	20.598%	19.710%	19.529%	19.301%	25.267%	26.621%	50.000%	
quarter	21.140%	20.855%	20.476%	19.807%	23.848%	25.153%	27.192%	50.000%

Statistics analysis

- 对于每个 test trial，统计 8个模型各自 hard trials 所交叠的 共有 hard trials~
- 具体做法如下：
  - 对于某个模型 M，以其 EER 下的阈值作为分水岭。对于 Target trials，选择出分数小于阈值的 trials；对于 Imposter trials，选择出分数大于 阈值的 trials。这些 trials 视为模型 M 的 hard trials。
- 分别选择出模型 M1-M8 各自的 hard trials；统计8组 hard trials 中所重叠的 trials。
- 结果如下：
  - SITW.Dev.Core: 338226 -> 6357
  - SITW.Eval.Core: 721788 -> 17486
  - VoxSRC_O_cl: 37611 -> 267
  - VoxSRC_E_cl: 579818 -> 4528
  - VoxSRC_H_cl: 550894 -> 7043

Performance of hard trials

PLDA EER （％）		VoxSRC			SITW	
	model_type	trials(％) (126:141)	trials_E(％) (2466:2062)	trials_H(％) (4277:2766)	Dev Core(％) (47:6310)	Eval Core(％) (97:17389)
i-vector	2048-400	99.21	99.96	99.98	97.87	98.97
	1024_400	99.21	99.96	99.98	97. 87	98.97
	512_200	99.21	99.96	99.98	97. 87	98.97
	256_200	99.21	99.96	99.98	97. 87	98.97
x-vector	double	99.21	99.96	99.98	97. 87	98.97
	raw	99.21	99.96	99.98	97. 87	98.97
	half	99.21	99.96	99.98	97. 87	98.97
	quarter	99.21	99.96	99.98	97. 87	98.97

Hard trials retrieval by SVM

- Based on these vectors, we can train an SVM and then check what trials are not correctly classified. These trials are hard trials.

	all	SVM hard	Threshold hard	tol	SVM是否包含了全部Threshold 的hard trials
SITW.Dev.Core	338226	18105	**6357**	2e-3	**×（6316）**
		17597	**6357**	2e-1	**×（6298）**
		18105	**6357**	1e-3	**×（6317）**
SITW.Eval.Core	721788	42643	**17486**	5e-1	**×（17040）**
VoxSRC_O_cl	37611	1399	**267**	1e-3	**√（267）**
VoxSRC_E_cl	579818	21390	**4528**	2e-3	**√（4528）**
VoxSRC_H_cl	550894	33397	**7043**	2e-2	**×（7042）**

Analysis

- 在五个测试列表上，使用阈值和SVM这两种方法选出的“hard trials”的重叠情况具有一致性。具体来说，使用threshold选出的“hard trials”，几乎完全包含了SVM的“hard trials”。
- Hard trials存在！

Human test

- What the properties of these hard trials?

- What do we humans sound like ?

Analysis

- 场景信息相关的数据分析:
- testcount >= 1 共4283组，target:nontarget=2034:2249
- 共答题数4316次

nontarget

	同场景	不同场景
数量	1005	1244
平均准确率	57.16%	65.85%

target

	同场景	不同场景
数量	656	1378
平均准确率	58.26%	52.94%

Conclusion

- Hard trials存在！
- 对于nontarget而言，场景不同的“hard trials”，平均准确率更高些；而target恰恰相反。说明不同场景条件的存在，会给human test提供有利的先验知识。