开源、鲁棒、可信的说话人识别

Open-source, Robust, Explainable Speaker Recognition

报 告 人：李 蓝 天
合作导师：郑方教授

清华大学 语音和语言技术中心
http://cslt.riit.tsinghua.edu.cn/

报告提要

- 说话人识别概述
- 开源：多场景中文明星数据库
- 鲁棒：跨、泛化、复杂场景
- 可信：模型可视化与性能评测

第一章节

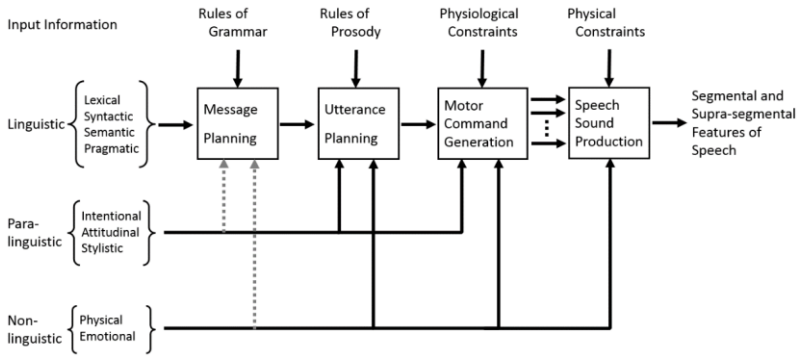
- 说话人识别概述
- 开源：多场景中文明星数据库
- 鲁棒：跨、泛化、复杂场景
- 可信：模型可视化与性能评测

基本概念

说话人识别，又称声纹识别

- 根据语音信号中表征说话人**个性**的**声纹**特征，利用计算机以及各种信息识别技术，自动地实现说话人身份辨识的一项生物特征识别技术。

视角一：语音信息编码



语言信息

文本内容
语言语义

副语言信息

话者声纹
意图态度

非语言信息

情绪情感
性别年龄

视角二：生物特征

声纹：具有 生理特性 的 行为特征

发展历史

阶段一 (1970~2000)

- 挑战：发音随机
- 技术：特征驱动
- 代表：MFCC/LPCC

阶段二 (1995~2010)

- 挑战：自由文本
- 技术：统计建模
- 代表：GMM-UBM

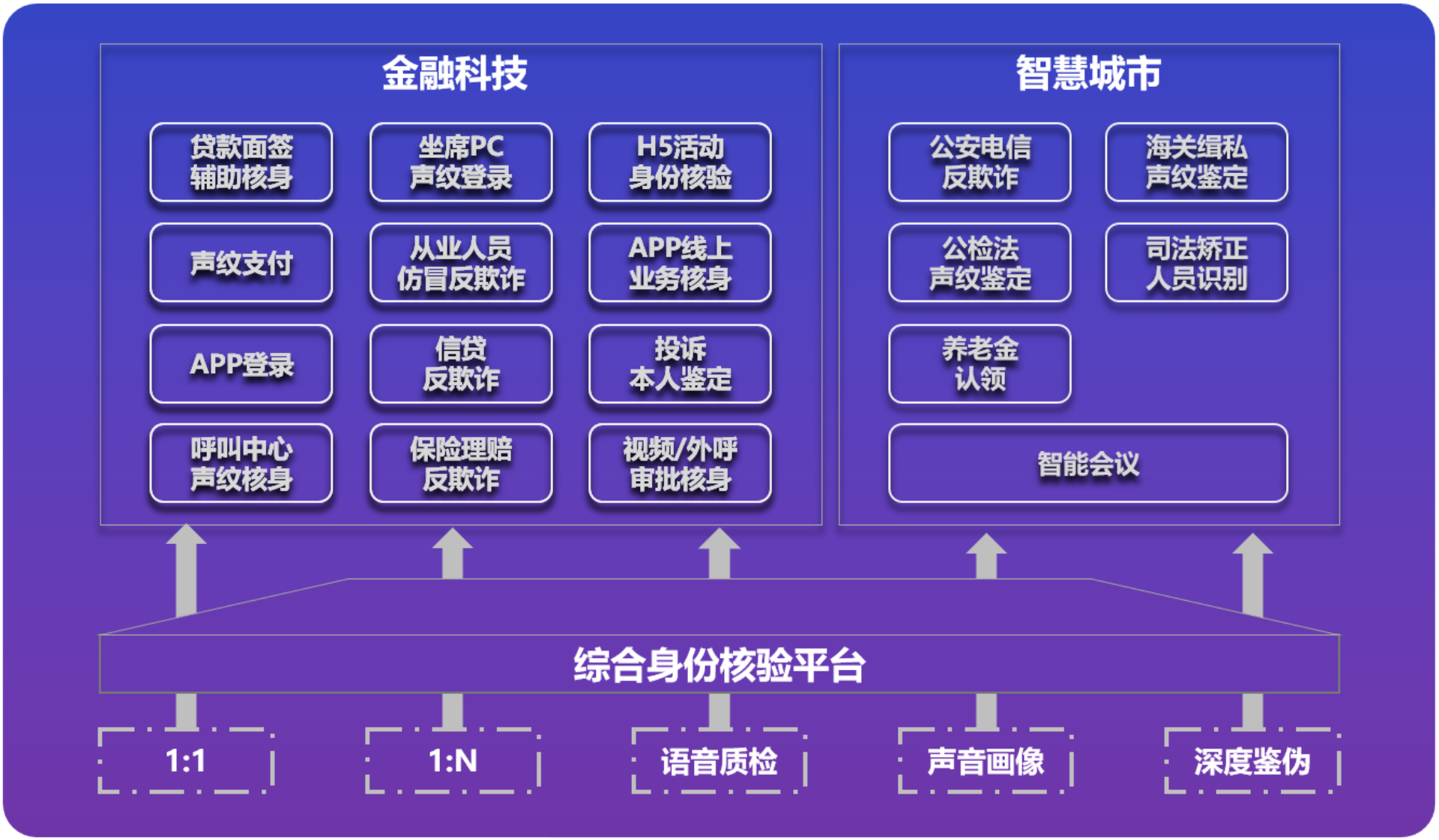
阶段三 (2005~2016)

- 挑战：会话扰动
- 技术：因子分析
- 代表：i-vector/PLDA

阶段四 (2017~)

- 挑战：复杂场景
- 技术：深度学习
- 代表：x-vector/E2E

应用场景


金融科技
贷款面签
辅助核身
坐席PC
声纹登录
H5活动
身份核验
声纹支付
从业人员
仿冒反欺诈
APP线上
业务核身
APP登录
信贷
反欺诈
投诉
本人鉴定
呼叫中心
声纹核身
保险理赔
反欺诈
视频/外呼
审批核身
智慧城市
公安电信
反欺诈
海关缉私
声纹鉴定
公检法
声纹鉴定
司法矫正
人员识别
养老金
认领
智能会议
综合身份核验平台
1:1
1:N
语音质检
声音画像
深度鉴伪

研究工作背景

问题一：说话人识别技术的性能上限

开源

- 当前主流的说话人评测集大都场景单一，难以描述多场景下的复杂性
- 单纯靠人工采集并标注一个大规模多场景数据集是极其费时费力的

问题二：多复杂场景下的鲁棒性

鲁棒

- 尽管当前说话人识别取得一定进展，但实际应用中的性能表现难言可靠
- 从应用视角出发，聚焦跨场景、泛化场景、复杂场景的三类鲁棒性问题

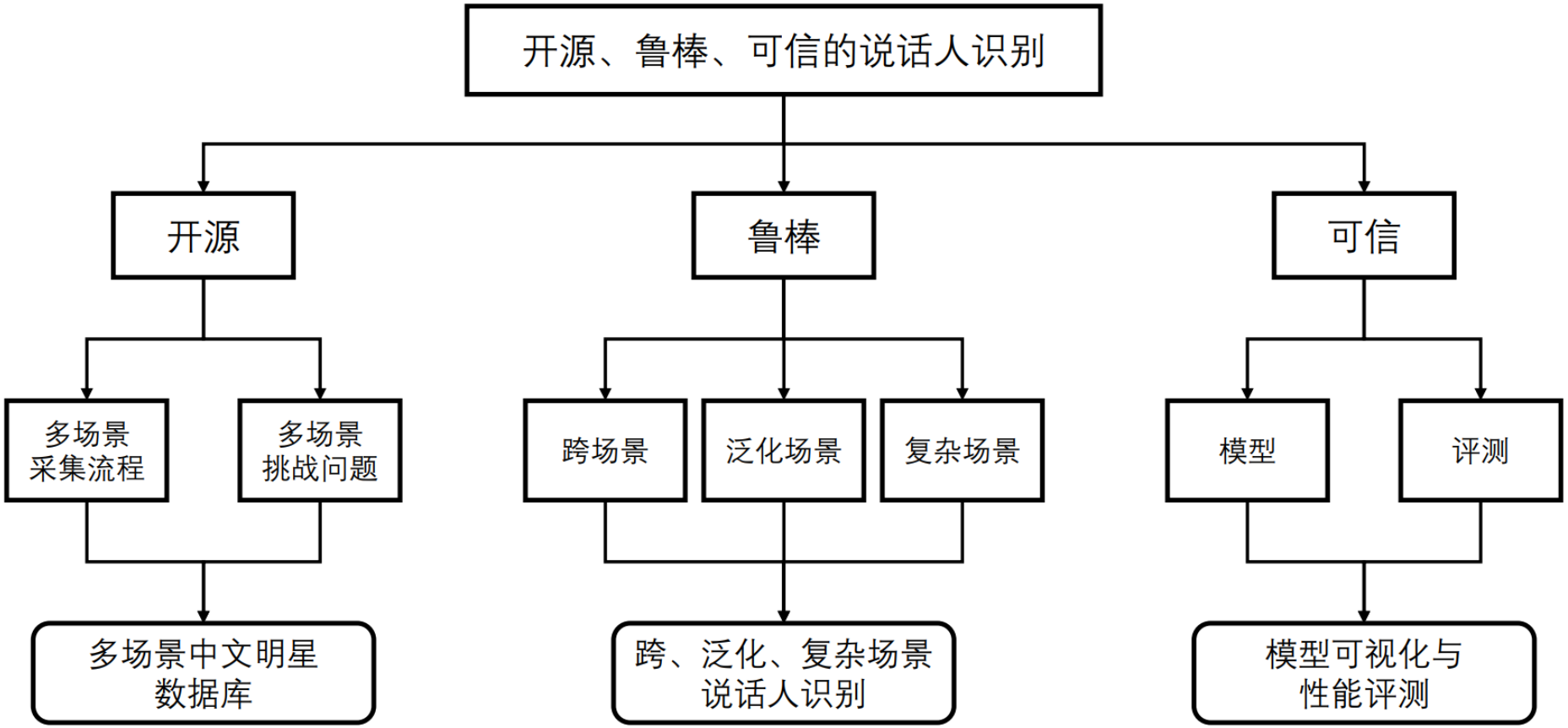
问题三：模型和评测的可解释性

可信

- 基于深度神经网络的说话人识别系统的黑盒属性，系统判决可信度存疑
- 基线评测与实际体验的性能不一致性，极大地困扰了研究者和从业者

研究工作框架

- 围绕前述三个研究问题，从数据、模型、应用出发，开展开源、鲁棒、可信的说话人识别研究



第二章节

- 说话人识别概述
- 开源：多场景中文明星数据库
- 鲁棒：跨、泛化、复杂场景
- 可信：模型可视化与性能评测

多场景的重要性

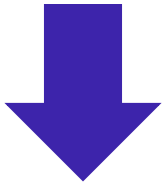
- 多场景：覆盖了各种变动性，是实际应用的真实挑战

时变	幼年 -> 青年 -> 中年 -> 老年
背景	户外 -> 课堂 -> 车载 -> 音乐
方式	哭声 -> 朗诵 -> 通话 -> 唱歌

- 在多场景下取得优秀的识别性能是说话人识别技术走向实际应用的**充要条件**

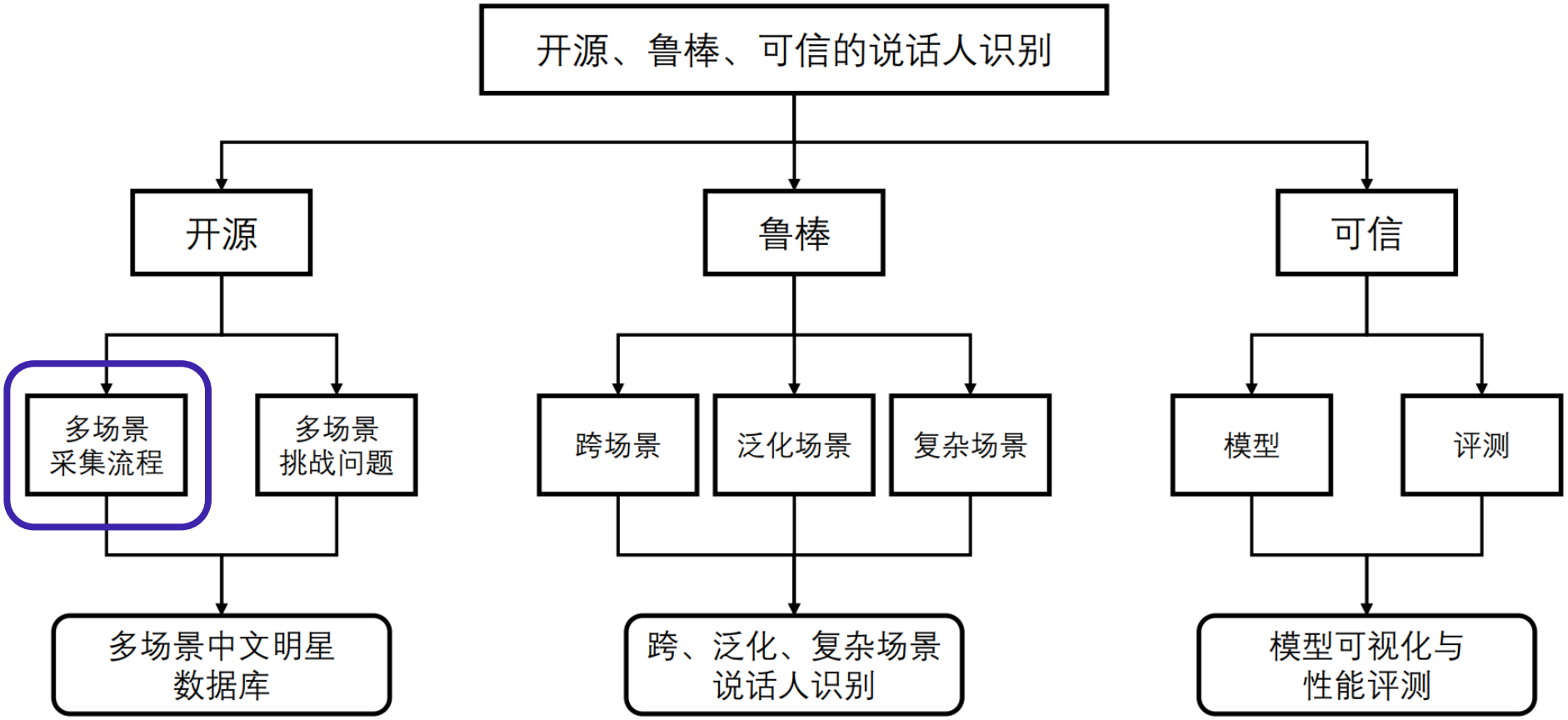
当前问题：如何解决多场景的数据空白

- 现有的说话人识别数据集大都是在**限定条件**下采集得到的，难以描述多场景下的复杂变动性
- **单纯人工**采集并标注一个大规模多场景数据集是极其费时费力的



- **研究工作一**：设计一套**自动化**多场景说话人数据采集工具，采集开源一套大规模多场景说话人识别数据集
- **研究工作二**：系统性分析多场景说话人识别的**挑战性**，初步探索可行的解决方法

研究工作一：多场景采集流程

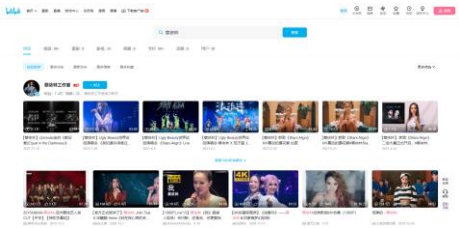


研究工作一：自动化采集流程

1 设计明星人名列表

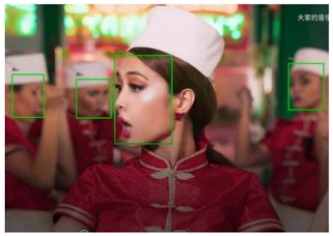
0001 周杰伦
0002 蔡依林
0003 刘德华
…

2 下载图片和视频

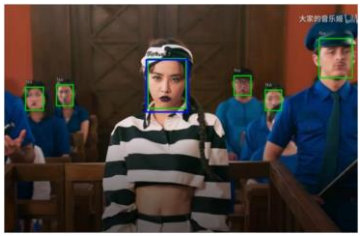


3 人脸检测与追踪

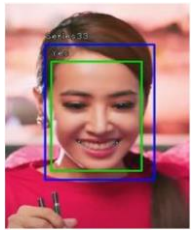
人脸检测
RetinaFace



人脸识别
ArcFace



人脸追踪
MOSSE

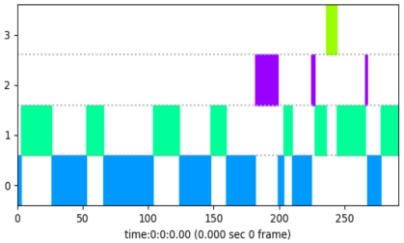


4 口唇同步检测

同步检测
SyncNet

5 说话人分割

话者分离
UIS-RNN



6 人工质检

要求：抽检正确率不低于90%

采集效率提升 4倍

多场景中文明星数据库 CN-Celeb

- 数据描述：3,000名中国明星，11种真实场景

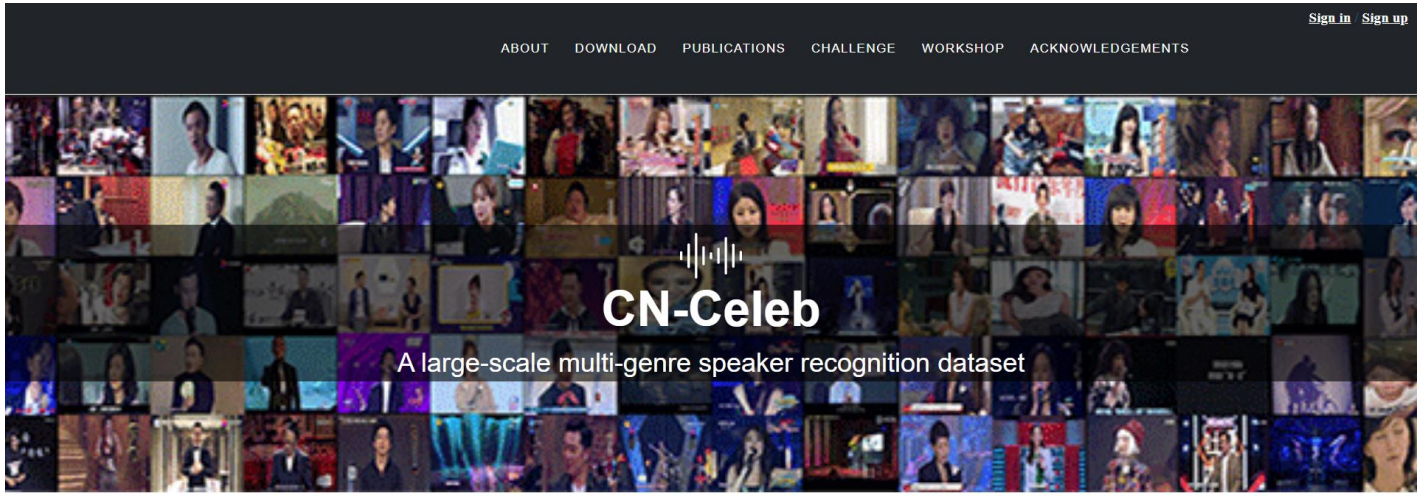
场景类型	CN-Celeb1			CN-Celeb2		
	说话人数	语音条数	语音时长 (h)	说话人数	语音条数	语音时长 (h)
广告 Advertisement	17	120	0.18	66	1,542	3.86
电视剧 Drama	160	7,247	6.43	268	13,116	16.32
娱乐 Entertainment	483	22,064	33.67	616	31,982	60.84
采访 Interview	780	59,317	135.77	519	34,024	81.28
直播 Live Broadcast	129	8,747	16.35	388	167,019	439.95
电影 Movie	62	2,749	2.20	133	4,449	5.77
戏剧 Play	69	4,245	4.95	127	14,992	22.04
朗诵 Recitation	41	2,747	4.98	218	58,231	129.18
唱歌 Singing	318	12,551	28.83	394	42,157	75.19
演讲 Speech	122	8,401	36.22	394	36,680	82.58
博客 Vlog	41	1,894	4.15	488	125,293	177.00
总计	1,000	130,109	273.73	2,000	529,485	1090.01

多场景中文明星数据库 CN-Celeb

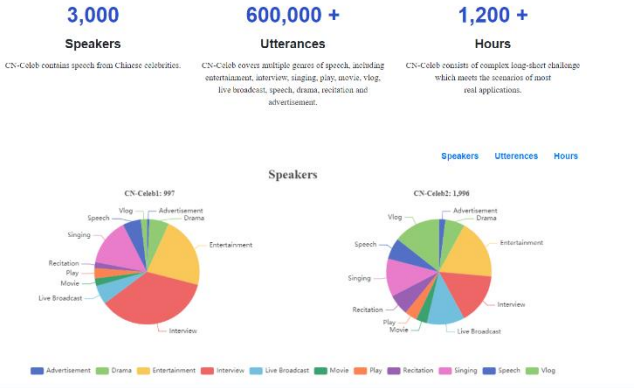
- 数据特色：多源化、多场景

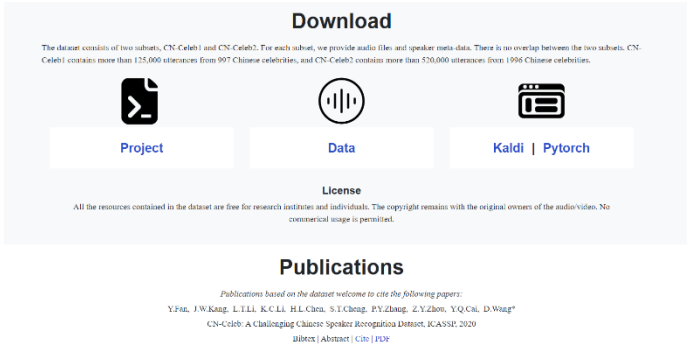
多场景中文明星数据库 CN-Celeb

- 数据开源: http://cnceleb.org/

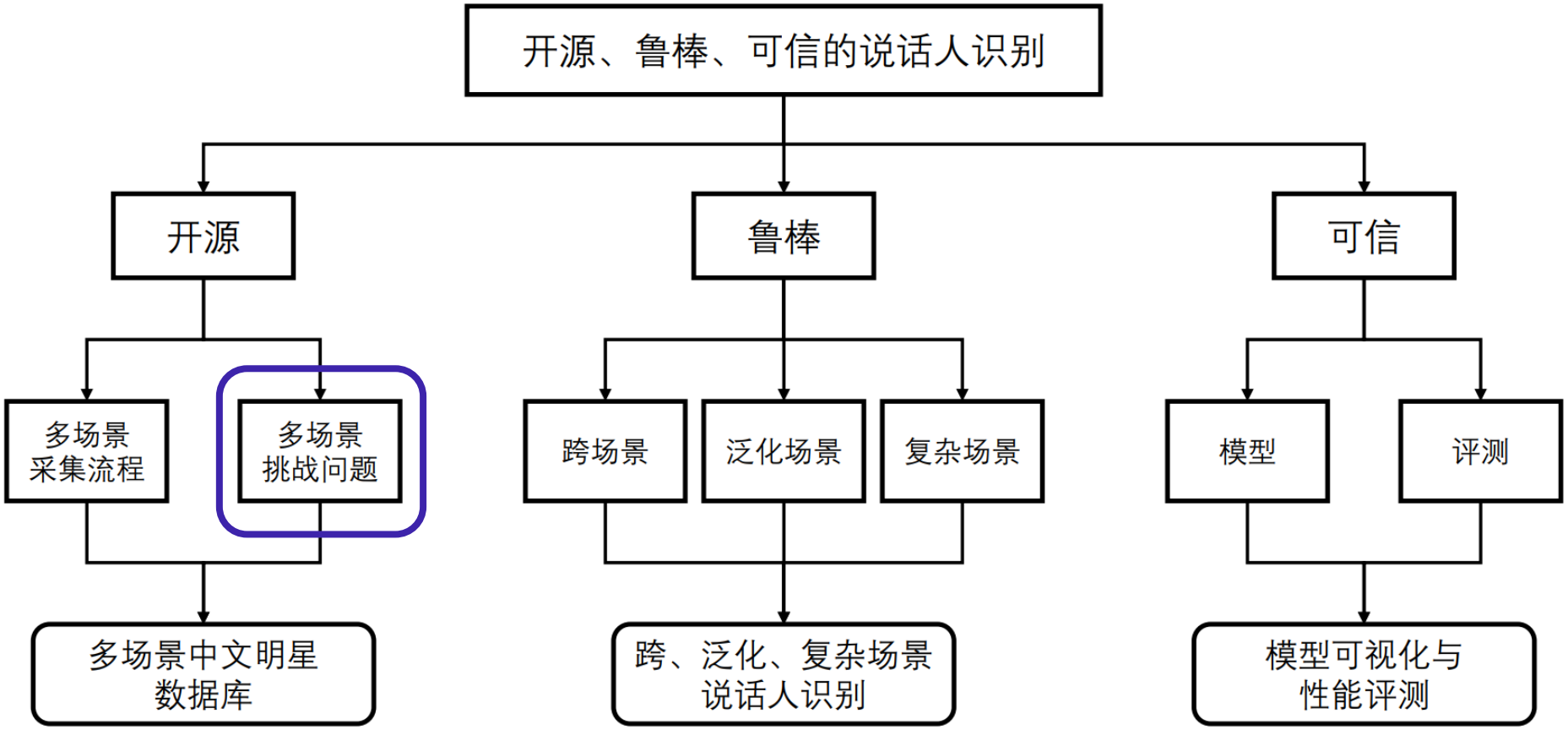


CN-Celeb is a multi-genre dataset covering 11 different genres in real world, collected from multiple Chinese open media sources.





研究工作二：多场景挑战性分析



研究工作二：多场景挑战性分析

基础实验

- i-vector 和 x-vector 基线系统 Kaldi

系统	训练集		测试集		
	前端	后端	SITW	SITW(S)	CN-Celeb.E
i-vector	VoxCeleb	VoxCeleb	5.66	7.41	18.37
x-vector	VoxCeleb	VoxCeleb	3.48	4.62	16.59

- 更强大的 x-vector 系统 TF-Kaldi

网络结构	池化方法	损失函数	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

主流技术无法解决多场景中的复杂性

研究工作二：多场景挑战性分析

- 多场景测试：以 x-vector 基线系统为例

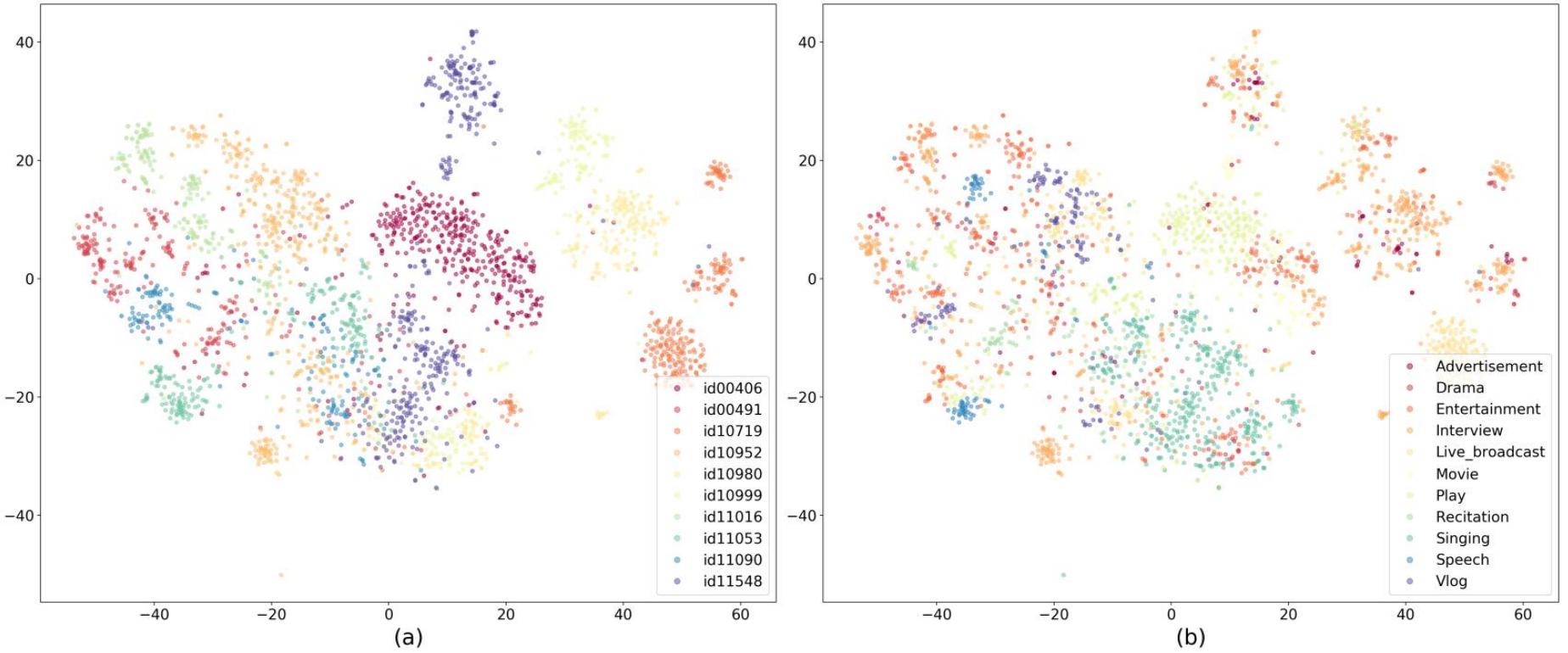
Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement	9.37	20.17	15.63	10.18	6.61	31.33	18.52	-	25.26	25.16	8.42	17.10
Drama	22.35	11.70	21.37	22.37	14.35	25.88	8.46	31.82	32.78	29.31	16.48	20.22
Entertainment	14.11	19.61	7.31	9.70	12.45	18.40	14.09	9.78	27.53	12.07	10.87	11.90
Interview	20.15	21.72	11.46	6.98	11.04	18.77	12.62	7.84	28.92	10.72	11.91	12.40
Live Broadcast	8.50	17.38	14.40	13.50	5.42	18.52	11.48	22.99	22.34	16.99	9.76	6.00
Movie	29.21	25.00	15.07	19.13	12.57	11.47	12.46	14.29	29.42	14.51	25.38	18.29
Play	9.09	16.84	14.91	19.29	15.09	24.67	11.56	5.88	28.25	11.82	18.18	12.65
Recitation	-	20.00	23.58	8.61	9.24	20.00	3.69	16.55	34.72	4.79	33.33	9.71
Singing	24.91	27.56	31.49	28.87	21.11	29.00	18.83	36.51	20.86	21.86	24.07	18.76
Speech	29.03	25.27	31.49	28.87	21.11	29.00	18.83	36.51	20.86	3.21	36.54	5.18
Vlog	5.26	19.73	31.49	28.87	21.11	29.00	18.83	36.51	20.86	36.63	5.31	5.32

多场景测试是极为挑战性的

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement					6.61							
Drama												
Entertainment			7.31									
Interview				6.98								
Live Broadcast					5.42							6.00
Movie												
Play								5.88				
Recitation							3.69			4.79		
Singing												
Speech								2.91		3.21		5.18
Vlog	5.26										5.31	5.32

研究工作二：多场景挑战性分析

可视化分析



- 多场景的复杂性使说话人类内分布变得复杂，类间分布存在明显交叠
- 多场景说话人识别是极具挑战性的

小结

开源：多场景中文明星数据库

- 设计了一套自动化数据采集平台，采集并开源了一套大规模多场景中文明星数据集 CN-Celeb。
- 系统性地分析了多场景说话人识别的挑战性。

第三章节

- 说话人识别概述
- 开源：多场景中文明星数据库
- 鲁棒：跨、泛化、复杂场景
- 可信：模型可视化与性能评测

从应用出发，定义三类鲁棒性问题

跨场景鲁棒性

- **特指注册-测试**场景失配问题
- 典型场景：跨信道、时变、远近场等

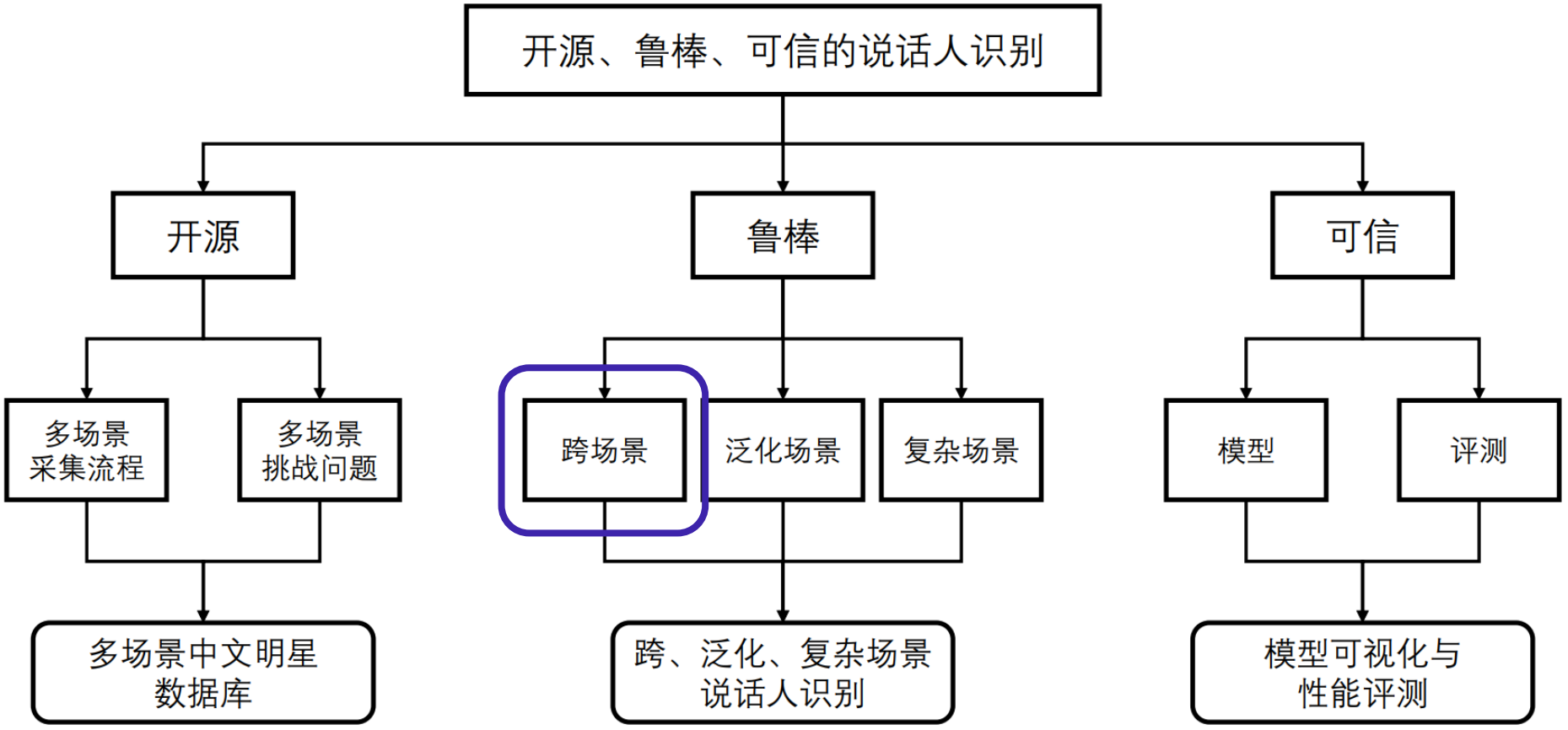
泛化场景鲁棒性

- **特指训练-部署**场景失配问题
- 典型场景：电话信道训练，网络信道部署

复杂场景鲁棒性

- 包括跨场景、泛化场景在内的综合问题
- 典型场景：困难场景测试、多场景测试

研究工作一：跨场景鲁棒性



研究工作一：跨场景鲁棒性

- 什么是跨场景问题？

跨信道场景

Enroll-Test	Baseline
AN-AN	0.797
AN-Mic	2.146
AN-iOS	1.425
Mic-AN	2.175
Mic-Mic	0.778
Mic-iOS	2.251
iOS-AN	1.599
iOS-Mic	2.216
iOS-iOS	0.920

时变场景

Enroll-Test	Baseline
1st-1st	4.799
1st-2nd	6.400
1st-3rd	6.863
1st-4th	6.884
1st-5th	7.108
1st-6th	7.856
1st-7th	7.906
1st-8th	7.881

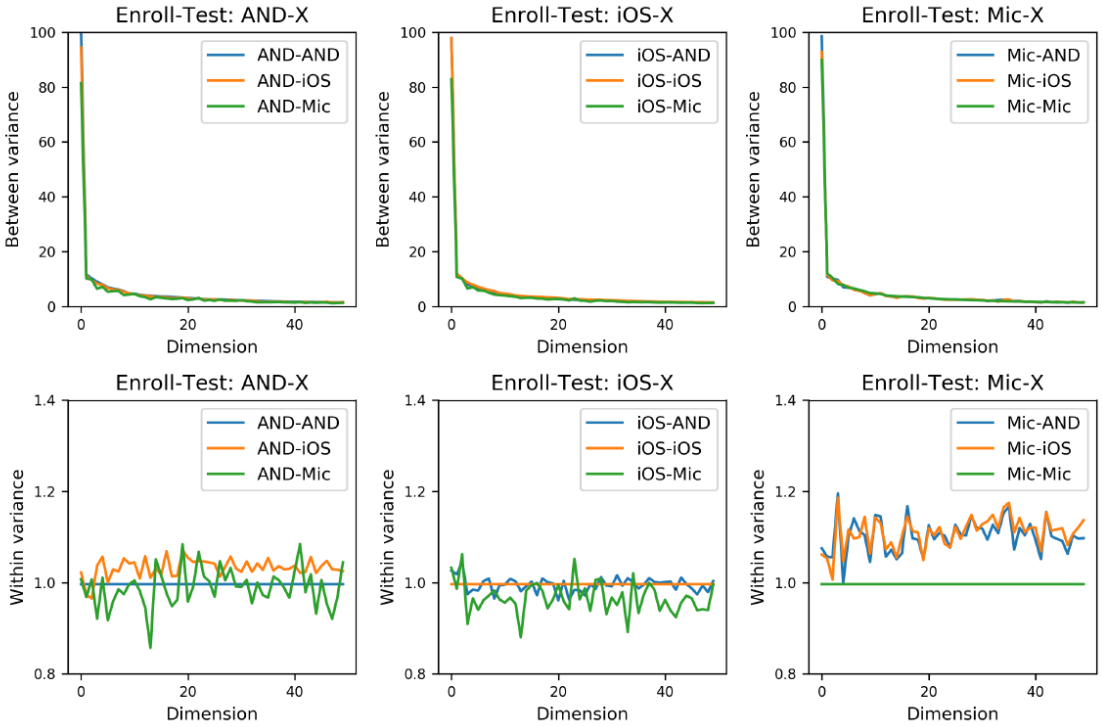
远近场场景

Enroll-Test	Baseline
1m-1m	0.620
1m-3m	3.968
1m-5m	4.866
3m-1m	1.938
3m-3m	0.891
3m-5m	3.244
5m-1m	3.566
5m-3m	2.834
5m-5m	1.135

当注册-测试失配时，性能下降的原因是什么？

注册-测试失配的根源

跨信道测试统计量分析



说话人类内和类间统计量

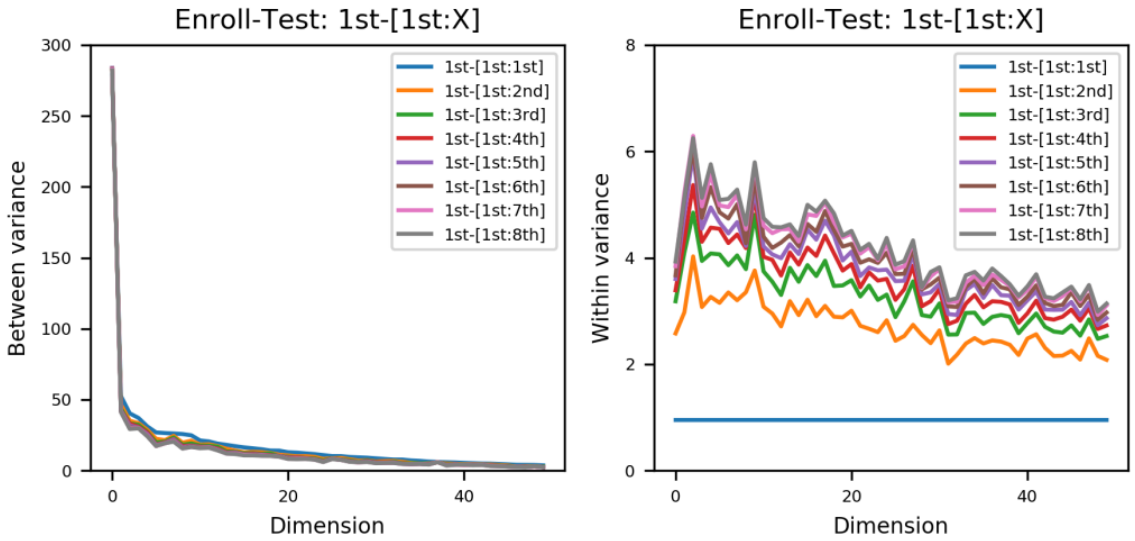
注册-测试	角度 (θ)	长度 (ℓ)
AND-iOS	0.940	4.673
iOS-Mic	5.199	10.789
Mic-AND	5.645	10.867

全局偏移量

- 说话人类内和类间方差相近
- 全局偏移现象显著

注册-测试失配的根源

时变测试统计量分析



说话人类内和类间统计量

注册-测试	角度 (θ)	长度 (ℓ)
1st-[1st:2nd]	0.085	1.615
1st-[1st:3rd]	0.159	2.326
1st-[1st:4th]	0.234	2.654
1st-[1st:5th]	0.269	2.991
1st-[1st:6th]	0.301	3.224
1st-[1st:7th]	0.346	3.439
1st-[1st:8th]	0.393	3.590

全局偏移量

- 说话人类间方差相近，全局偏移较小
- 说话人类内方差随着时变累积，逐渐变大

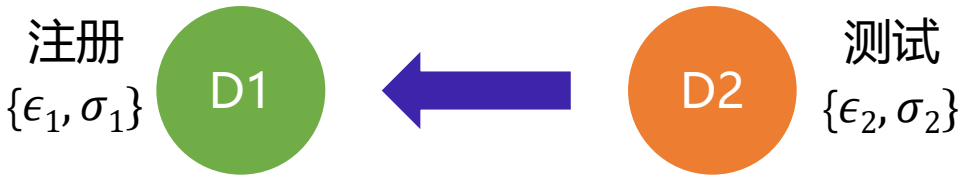
注册-测试失配的根源

注册场景和测试场景之间的统计量不一致性

可行的解决方案

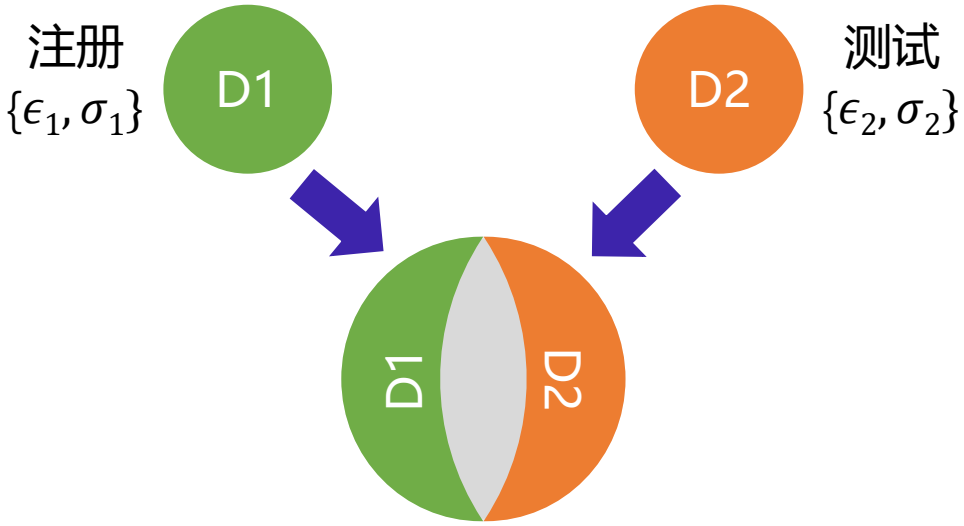
二者均非理论最优

场景自适应



- 测试数据映射到注册场景中；使用注册场景的统计模型进行打分
- 然而，测试数据使用测试场景的统计模型打分才是理论最优的

混合场景训练



- 混合注册数据和测试数据，训练一个二者共享的统计模型
- 等价于对注册场景统计模型和测试场景统计模型的等权插值

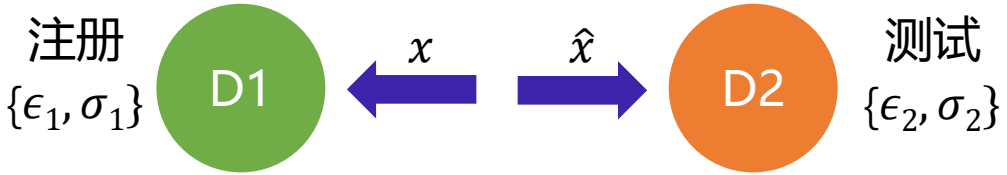
从对数似然比到统计量解耦

对数似然比的解耦形式

$$LLR(x|u_k) = \log p(x|H0) - \log p(x|H1)$$
$$= \log p_k(x) - \log p(x)$$
$$= \log \int p(u_k|x_1^k, \dots, x_n^k)\, p(x|u_k) du - \log p(x)$$

- 注册项：$p(u_k|x_1^k, \dots, x_n^k)$
- 预测项：$p_k(x) = \int p(u_k|x_1^k, \dots, x_n^k)\, p(x|u_k) du$
- 归一项：$p(x)$

当注册数据和测试数据使用各自最优的统计模型，方可得到理论最优的打分



基于线性映射的统计量解耦

- 针对预测项失配问题，建立注册和测试之间的对应关系
  - 线性映射：$x = M\hat{x} + b$
  - 预测项：$p_k(\hat{x}; M, b) = p\left(M\hat{x} + b \middle| x_1^k, \dots, x_n^k\right)$

$$= \int p(M\hat{x} + b|u_k) p\left(u_k \middle| x_1^k, \dots, x_n^k\right) du_k$$

$$= N\left(M\hat{x} + b; \frac{\epsilon_1\sigma_1}{n\epsilon_1 + \sigma_1}\bar{x}_k, \left(\sigma + \frac{\epsilon_1\sigma_1}{n\epsilon_1 + \sigma_1}\right) I\right)$$

- 优化目标：最大似然估计

$$\mathcal{L}(M, b) = \sum_{k=1}^{K}\sum_{i=1}^{N} \log p_k(\hat{x}_{ik}; M, b)$$

- 注册-测试失配场景下的打分形式

$$LLR(\hat{x}|k) \propto -\frac{1}{\sigma_1 + \frac{\epsilon_1\sigma_1}{n\epsilon_1 + \sigma_1}} \left|\left|M\hat{x} + b - \tilde{u}_k\right|\right|^2 + \frac{1}{\epsilon_2 + \sigma_2} \left|\left|\hat{x}\right|\right|^2$$

实验一：跨信道

- 统计量解耦方法优于混合训练和场景自适应方法

注册-测试	Base	Base + AS-norm	NPD			PD			
			IDVC	GSC	WVA	MCT	CAT	SD/LT	SD/LT + AS-norm
AND-AND	0.797	0.684	-	-	-	-	-	-	-
AND-Mic	2.146	1.316	1.768	1.764	2.165	1.151	1.245	0.981	**0.943**
AND-iOS	1.425	0.741	1.354	1.382	1.401	1.161	1.312	**0.623**	0.760
Mic-AND	2.175	1.382	1.665	1.665	2.033	1.161	1.189	**0.712**	1.128
Mic-Mic	0.778	0.849	-	-	-	-	-	-	-
Mic-iOS	2.251	1.231	1.920	1.892	2.081	1.293	1.481	**0.812**	1.038
iOS-AND	1.599	0.972	1.382	1.430	1.590	1.156	1.184	**0.755**	0.854
iOS-Mic	2.216	1.330	1.726	1.759	2.231	1.137	1.231	1.052	**0.948**
iOS-iOS	0.920	0.684	-	-	-	-	-	-	-

MCT：混合场景训练；CAT：场景自适应；SD/LT：统计量解耦

实验二：时变测试

- 统计量解耦方法优于混合训练和场景自适应方法

注册-测试	Base	Base + AS-norm	NPD			PD			
			IDVC	GSC	WVA	MCT	CAT	SD/LT	SD/LT + AS-norm
1st-[1st:1st]	4.799	5.016	-	-	-	-	-	-	-
1st-[1st:2nd]	6.400	6.306	6.258	6.346	5.934	5.258	5.549	**4.339**	5.218
1st-[1st:3rd]	6.863	6.282	6.669	6.777	6.156	4.976	5.233	**4.062**	5.111
1st-[1st:4th]	6.884	6.256	6.850	6.810	6.084	4.619	4.882	**3.710**	4.612
1st-[1st:5th]	7.108	6.309	6.846	7.022	6.230	4.348	4.804	**3.678**	4.413
1st-[1st:6th]	7.856	6.832	7.595	7.768	6.938	4.348	4.861	**3.661**	4.298
1st-[1st:7th]	7.906	7.015	7.618	7.825	7.005	4.262	4.903	**3.749**	4.473
1st-[1st:8th]	7.881	7.034	7.737	7.815	6.993	4.300	5.041	**3.937**	4.499

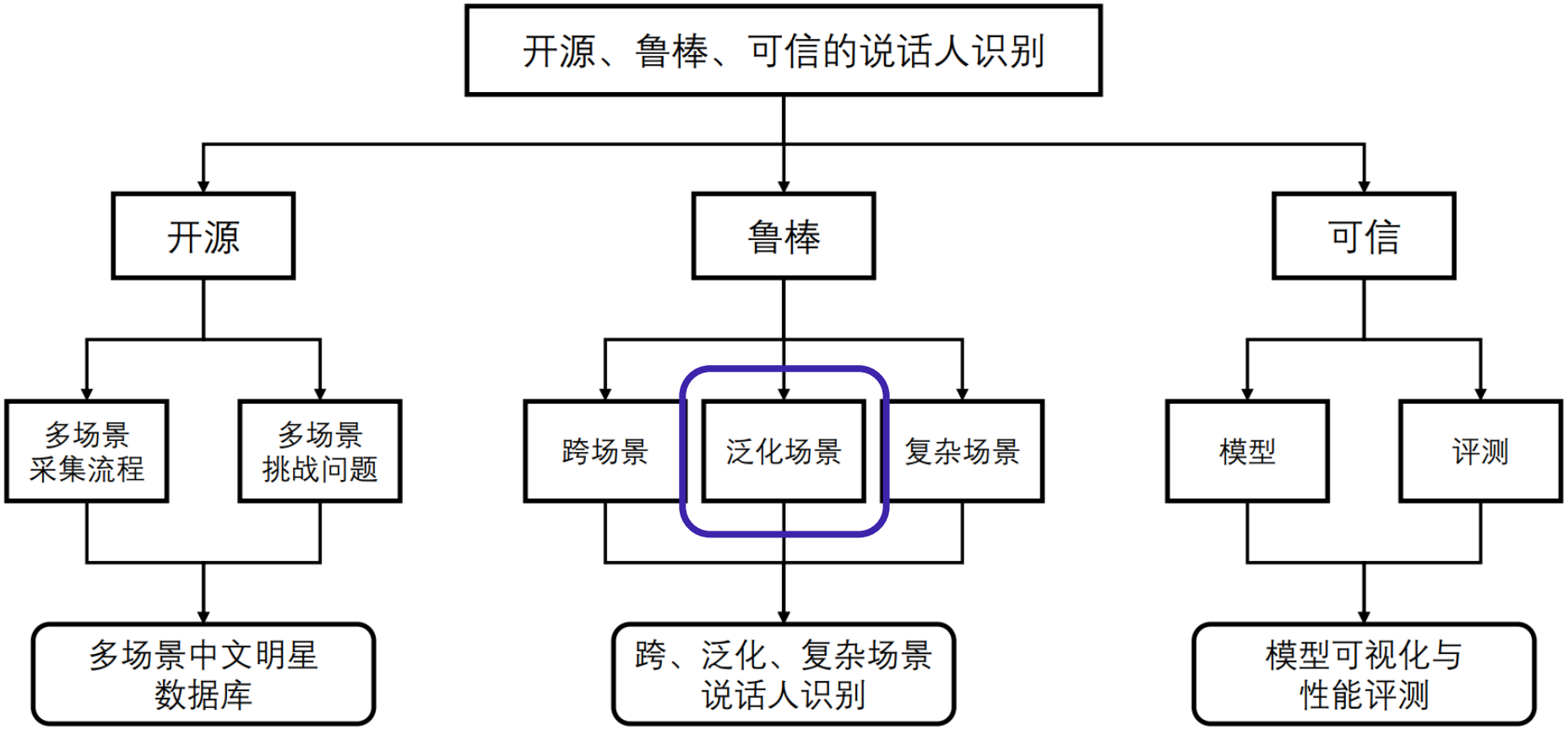
MCT：混合场景训练；CAT：场景自适应；SD/LT：统计量解耦

小结

跨场景鲁棒性

- 证明了注册-测试场景失配的根源是统计量不一致问题。
- 提出了统计量解耦的打分形式，将打分过程进行分解，并在不同阶段使用最优的统计量，从而得到理论最优打分。
- 通过两类典型跨场景测试，验证了该方法的理论性和有效性。

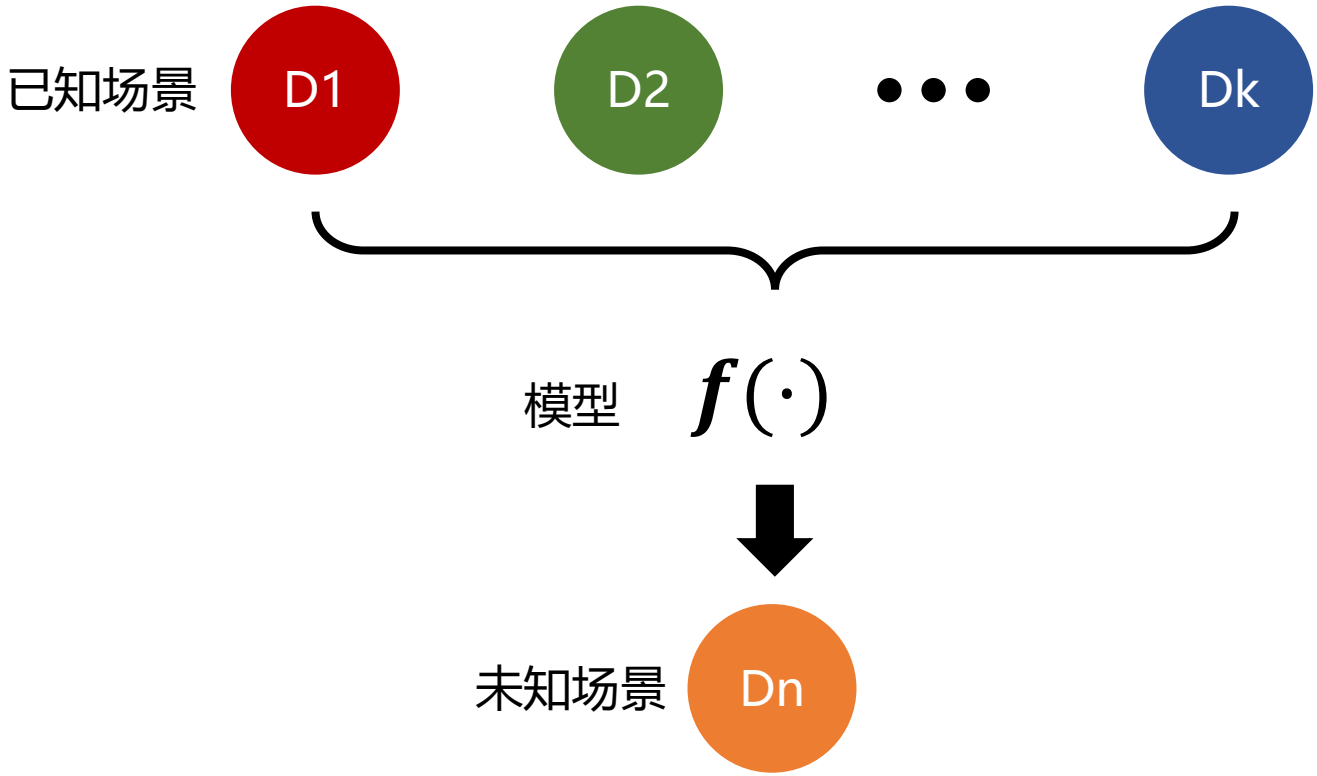
研究工作二：泛化场景鲁棒性



研究工作二：泛化场景鲁棒性

什么是泛化场景问题？

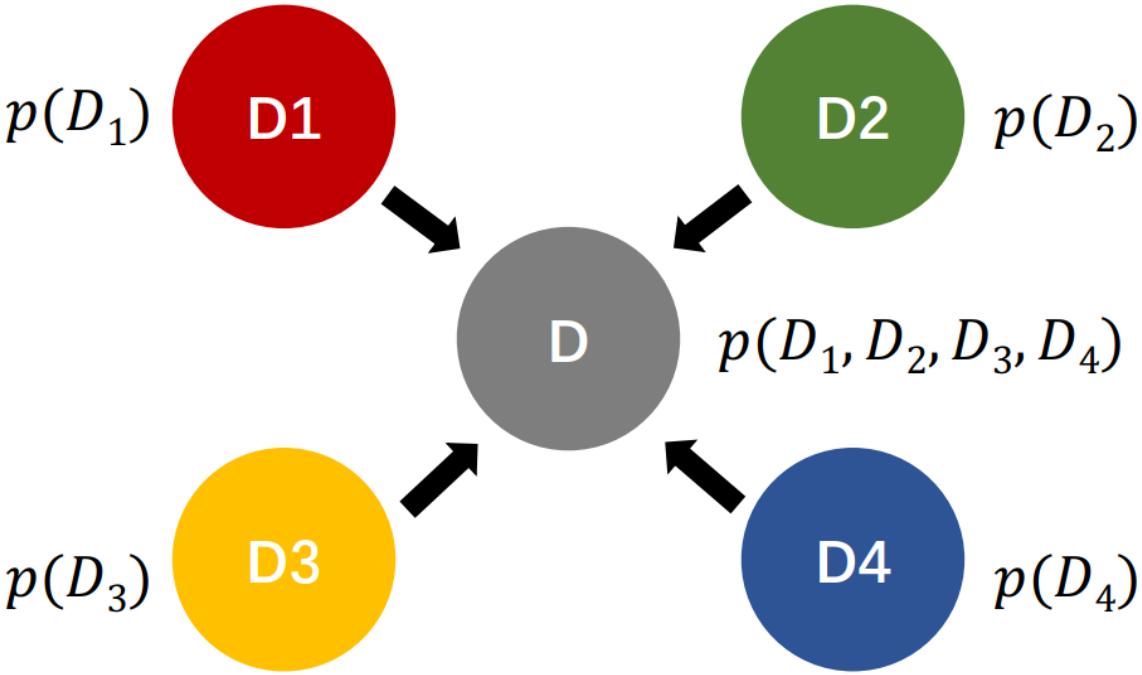
- 已知K类场景，优化预测模型 $f(\cdot)$，使之在未知场景上取得好的表现



优化目标：学习一个场景无关的说话人空间

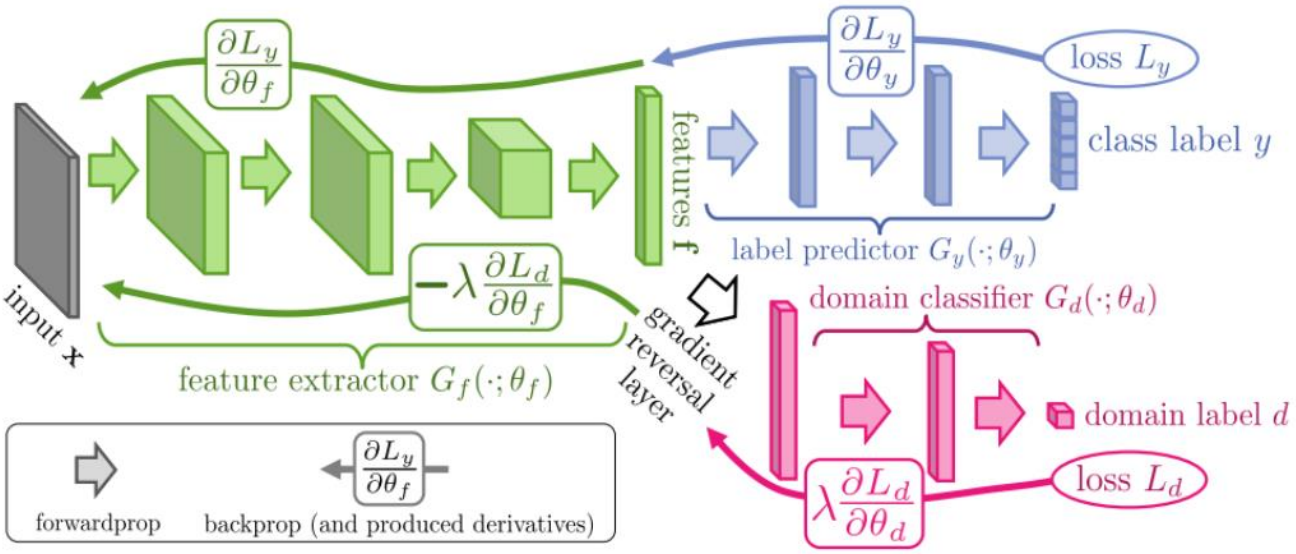
主流方法一：混合场景训练

- 基本思想：混合多个场景数据，优化训练目标；强化说话人信息，削弱场景扰动，学习场景无关的表征空间



主流方法二：场景对抗学习

- 基本思想：引入梯度反转层，将说话人信息和场景扰动进行分离，学习场景无关的表征空间



$$\mathcal{L} = \mathcal{L}_y - \lambda\mathcal{L}_d$$

- 与混合场景训练相比，增加了场景梯度反转

所提出方法：场景鲁棒训练

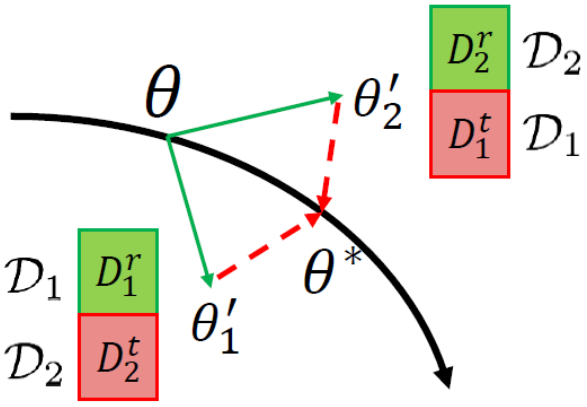
- 基本思想：结合模型无关元学习和鲁棒性优化算法，学习场景无关的表征空间
  - 数据准备：每一个 batch 数据取自两个不同场景 D_1 和 D_2，分别用于局部更新和全局更新。
  - 第一步：采样 $m_1^r \sim D_1$，进行局部更新

    $$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}(f_\theta; m_1^r)$$

  - 第二步：采样 $m_2^t \sim D_2$，进行全局更新

    $$\mathcal{L}(f_{\theta'}; m_2^t) = \mathcal{L}(f_{\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta; m_1^r)}; m_2^t)$$

    $$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}(f_{\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta; m_1^r)}; m_2^t)$$



所提出方法：场景鲁棒训练

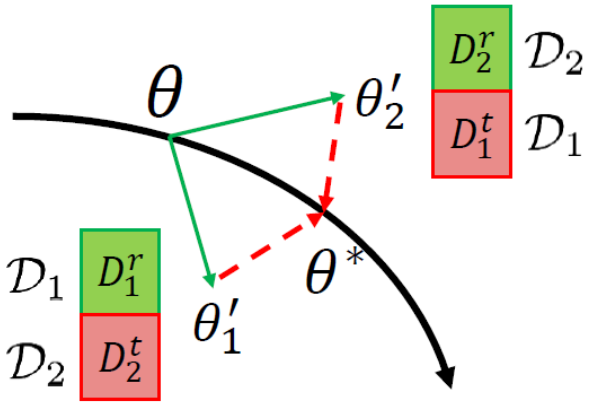
理论分析：以两个场景 D_1 和 D_2 为例

- 全局更新

$$\ell(\theta';m_2^t)=\ell(\theta-\alpha\nabla_\theta\ell(\theta;m_1^r);m_2^t)$$

$$\ell(\theta';m_1^t)=\ell(\theta-\alpha\nabla_\theta\ell(\theta;m_2^r);m_1^t)$$

$$\ell(\theta';m^t)=\ell(\theta';m_1^t)+\ell(\theta';m_2^t)$$



- 泰勒展开

$$\ell(\theta';m^t)=\ell(\theta;m_1^t)+\ell(\theta;m_2^t)-2*\alpha\langle\nabla_\theta\ell(\theta;m_1^r),\nabla_\theta\ell(\theta;m_2^t)\rangle$$

混合场景训练　　场景梯度趋同

- 通过使用两个场景数据分步更新，鼓励不同场景的梯度方向趋同，促使模型学习出场景无关的表征空间

实验结果

x-vector 系统

Method	Seen domain		Unseen domain				
	Cellular	Landline	Satellite	Voip	Android	IOS	Microphone
MDT	4.58	4.06	10.31	4.13	0.46	0.38	0.70
DAT	4.44	3.96	10.23	4.08	0.45	0.40	0.67
DRT	**4.24**	**3.53**	**9.42**	**3.89**	**0.32**	**0.27**	**0.54**

i-vector 系统

Method	Seen domain		Unseen domain				
	Cellular	Landline	Satellite	Voip	Android	IOS	Microphone
MDT	5.05	5.16	9.60	4.85	1.73	0.85	3.07
DAT	4.95	5.20	9.83	4.74	1.73	0.71	3.09
DRT	**3.72**	**3.83**	**8.15**	**3.24**	**0.87**	**0.47**	**1.44**

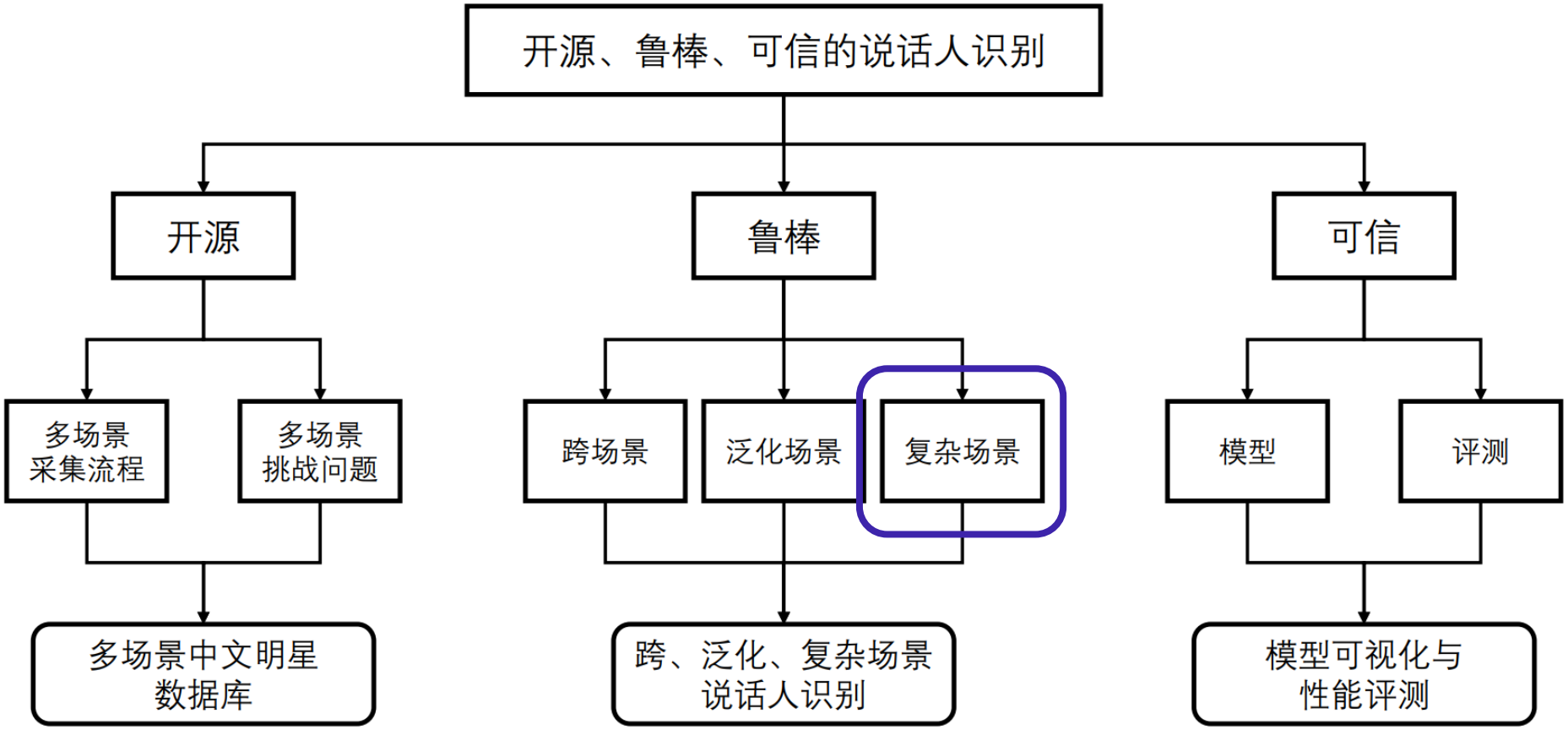
MDT：混合场景训练；DAT：场景对抗训练；DRT：场景鲁棒训练

小结

泛化场景鲁棒性

- 通过借鉴模型无关元学习的两步训练策略以及鲁棒性优化的学习机制，提出了场景鲁棒训练方法，学习场景无关的表征空间。
- 对比主流混合场景训练和场景对抗训练，本方法具有一致性优势。

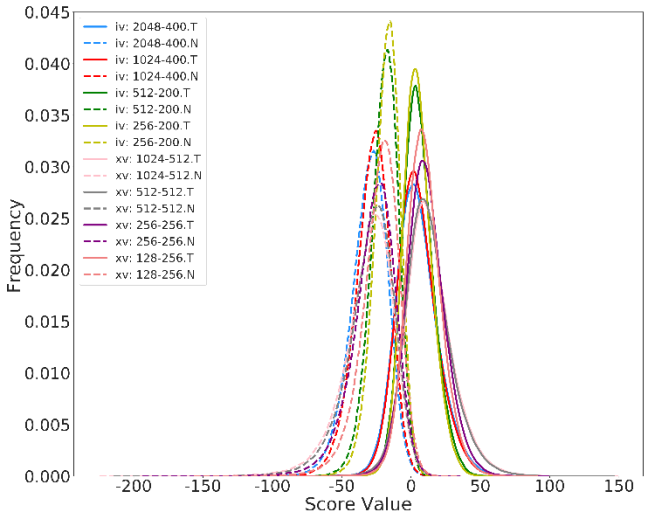
研究工作三：复杂场景鲁棒性

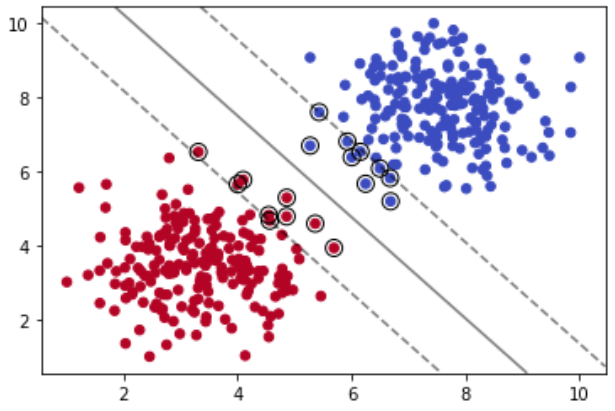


研究工作三：复杂场景鲁棒性

- 目标对象：多场景和难场景
  - 多场景：CN-Celeb

  - 难场景：各种不同模型均难以正确判别

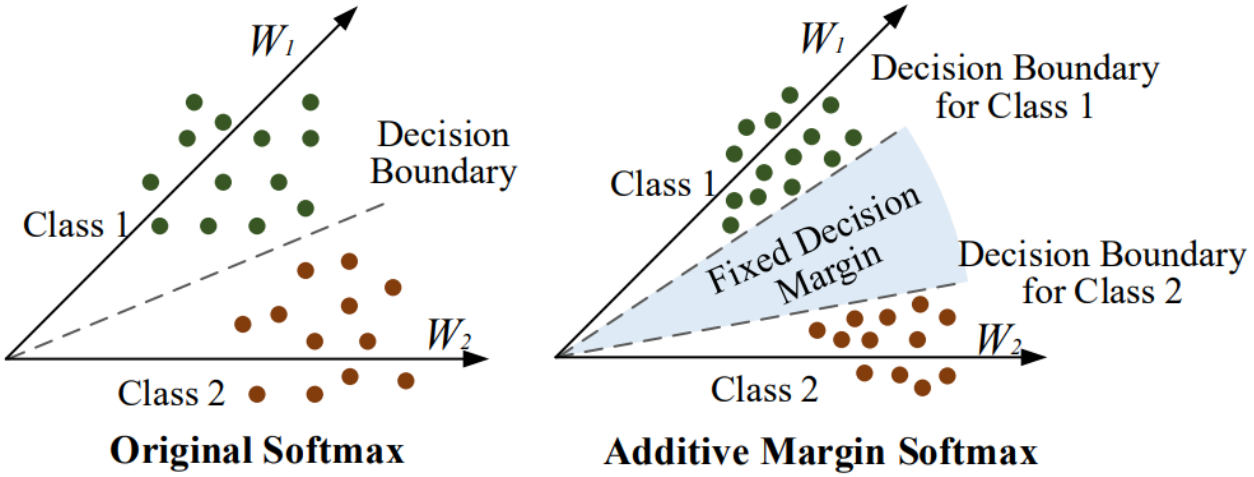




主流方法

基于边界 Softmax 的损失函数

- 在标准 Softmax 基础上，通过在目标 logit 中引入一个固定的边界量 m，来增大目标 logit 和 非目标 logit 之间的边界



以 AM-Softmax 为例

- AM-Softmax: Additive Margin Softmax
- 在目标 logit 中引入边界量 m

$$\mathcal{L}_{\text{AM-Softmax}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i,i})-m)}}{e^{s(\cos(\theta_{y_i,i})-m)}+\sum_{j\neq y_i}e^{s(\cos(\theta_{j,i}))}}$$

- 它的假定是相比于非目标 logit，损失函数将更关注于目标 logit，从而能更好地划分目标类和非目标类。

?

边界量 m 并不会增大边界

- 重写 AM-Softmax 损失函数

$$
\begin{aligned}
\mathcal{L}_{\text{AM-Softmax}} &= \frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i,i})-m)}+\sum_{j\neq y_i}e^{s(\cos(\theta_{j,i}))}}{e^{s(\cos(\theta_{y_i,i})-m)}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\Big\{1+\sum_{j\neq y_i}e^{-s(\cos(\theta_{y_i,i})-\cos(\theta_{j,i})-m)}\Big\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\Big\{1+e^{sm}\sum_{j\neq y_i}e^{-s(\cos(\theta_{y_i,i})-\cos(\theta_{j,i}))}\Big\}
\end{aligned}
$$

- 当 s=1、m=0 时，AM-Softmax 回退成 Softmax
- m 仅仅是改变了损失函数的样式，而并不会增大目标 logit 和 非目标 logit 之间的边界

特殊情况

- 简单样本：目标 logit 占主导

$$\frac{e^{(\cos(\theta_{y_i,i})-m)}}{e^{(\cos(\theta_{y_i,i})-m)}+\sum_{j\neq y_i}e^{(\cos(\theta_{j,i}))}}\approx 1$$

增大 m 会突出简单样本

$$\log\{1+e^{m}\sum_{j\neq y_i}e^{-(\cos(\theta_{y_i,i})-\cos(\theta_{j,i}))}\}\approx e^{m}\sum_{j\neq y_i}e^{-(\cos(\theta_{y_i,i})-\cos(\theta_{j,i}))}$$

- 困难样本：目标 logit 变弱

$$\frac{e^{(\cos(\theta_{y_i,i})-m)}}{e^{(\cos(\theta_{y_i,i})-m)}+\sum_{j\neq y_i}e^{s(\cos(\theta_{j,i}))}}\ll 1$$

改变 m 对困难样本无效

$$\log\{1+e^{m}\sum_{j\neq y_i}e^{-(\cos(\theta_{y_i,i})-\cos(\theta_{j,i}))}\}\approx m+\log\sum_{j\neq y_i}e^{-(\cos(\theta_{y_i,i})-\cos(\theta_{j,i}))}$$

- 简单样本权重大于困难样本，存在困难场景鲁棒性风险

在 AM-Softmax 引入真正的边界

- 标准的最大边界训练 (Max-margin training)

$$\mathcal{L}_{\text{margin}} = \max(0, d_p - d_n + m)$$

- AM-Softmax 中缺失了 max 操作

- 改进方法：在 AM-Softmax 中引入 max 操作，修正得到真正的 AM-Softmax (*Real* AM-Softmax)

$$\mathcal{L}_{\text{RAM-Softmax}} = \frac{1}{N}\sum_{i=1}^{N}\log\Big\{1+\sum_{j\neq y_i} e^{\max\{0,-s(\cos(\theta_{y_i,i})-\cos(\theta_{j,i})-m)\}}\Big\}$$

- Real AM-Softmax 鼓励模型聚焦在困难非目标 logits 上，而忽略简单非目标 logits

基础实验

在 VoxCeleb1 和 SITW 上的 EER(%)

Objective	Hyperparameters	VoxCeleb1	VoxCeleb1-H	VoxCeleb1-E	SITW.Dev.Core	SITW.Eval.Core
AM-Softmax	m = 0.20, s = 30	**1.739**	2.895	1.724	**2.811**	3.362
Real AM-Softmax	m = 0.20, s = 30	1.872	3.068	1.883	3.466	3.718
	m = 0.25, s = 30	1.819	2.914	1.781	3.350	3.554
	m = 0.30, s = 30	1.755	**2.812**	**1.696**	3.003	3.417
	m = 0.35, s = 30	1.808	2.888	1.747	2.849	**3.335**

- Real AM-Softmax 一致性优于 AM-Softmax

消融测试

Objective	Hyperparameters	VoxCeleb1-H	VoxCeleb1-E	SITW.Eval.Core
AM-Softmax	s = 1, m = 0.00	17.955	10.916	11.400
	s = 1, m = 0.20	18.197	11.053	11.372
	s = 30, m = 0.20	2.895	1.724	3.362
Real AM-Softmax	s = 1, m = 0.00	18.650	11.908	12.657
	s = 1, m = 0.30	13.386	7.539	8.584
	s = 30, m = 0.30	2.812	1.696	3.417

- m 对 AM-Softmax 无效；而对 Real AM-Softmax 至关重要

复杂场景测试

难场景测试

损失函数	超参数	VoxCeleb1-H	VoxCeleb1-E	SITW.Eval.Core
AM-Softmax	s = 30, m = 0.20	39.794	38.970	36.082
Real AM-Softmax	s = 30, m = 0.25	39.899	37.814	35.052
	s = 30, m = 0.30	**39.175**	36.861	36.082
	s = 30, m = 0.35	39.749	**36.821**	**32.990**

多场景测试

损失函数	超参数	CN-Celeb.Eval
AM-Softmax	m = 0.10, s = 30	11.450
Real AM-Softmax	m = 0.10, s = 30	11.618
	m = 0.15, s = 30	11.323
	m = 0.20, s = 30	**11.049**
	m = 0.25, s = 30	11.422

- 相比 AM-Softmax，Real AM-Softmax 在多场景、难场景上表现突出

小结

复杂场景鲁棒性

- 针对复杂场景中的多场景、难场景问题，分析了主流 AM-Softmax 无法满足最大边界的假设，提出了真正边界的 AM-Softmax。
- 实验表明该方法在多场景和难场景中具有更好的鲁棒性。

第四章节

- 说话人识别概述
- 开源：多场景中文明星数据库
- 鲁棒：跨、泛化、复杂场景
- 可信：模型可视化与性能评测

课题起因

中国科学：信息科学　2020 年　第 50 卷　第 9 期：1281–1302
SCIENTIA SINICA Informationis
纪念《中国科学》创刊 70 周年专刊 · 评述
《中国科学》杂志社
SCIENCE CHINA PRESS

迈向第三代人工智能

张钹*，朱军，苏航

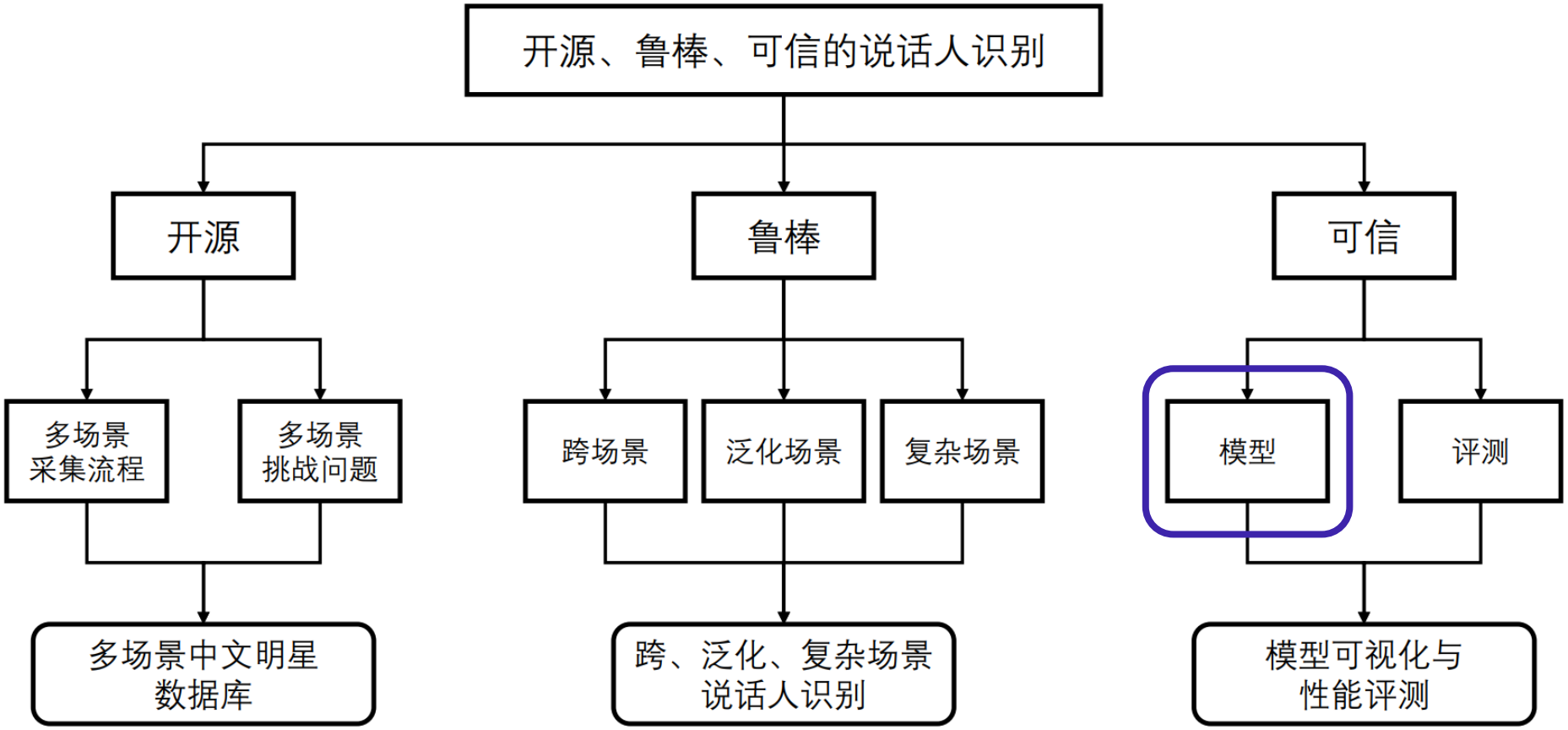
清华大学人工智能研究院，北京 100084
* 通信作者. E-mail: dcszb@tsinghua.edu.cn

深度学习为何如此脆弱，这样容易受攻击，被欺骗和不安全. 原因只能从机器学习理论本身去寻找. 机器学习的成功与否与 3 项假设密切相关，由于观察与测量数据的不确定性，所获取的数据一定不完备和含有噪声，这种情况下，神经网络结构 (备选函数族) 的选择极为重要，如果网络过于简单，则存在欠拟合 (under-fitting) 风险，如果网络结构过于复杂，则出现过拟合 (overfitting) 现象. 虽然通过各种正则化的手段，一定程度上可以降低过拟合的风险，但是如果数据的质量差，则必然会导致推广能力的严重下降. 此外，深度学习的 “黑箱” 性质是造成深度学习推广能力差的另一个原因，以图像识别为例，通过深度学习只能发现重复出现的局部片段 (模式)，很难发现具有语义的部件. 文献 [33] 描

纪初进入高潮，大有替代符号主义之势. 今天看来，这两种范式只是从不同的侧面模拟人类的心智 (或大脑)，具有各自的片面性，依靠单个范式不可能触及人类真正的智能. 需要建立新的可解释和鲁棒的 AI 理论与方法，发展安全、可信、可靠和可扩展的 AI 技术. 为实现这个目标，需要将这两种范式结合

可信

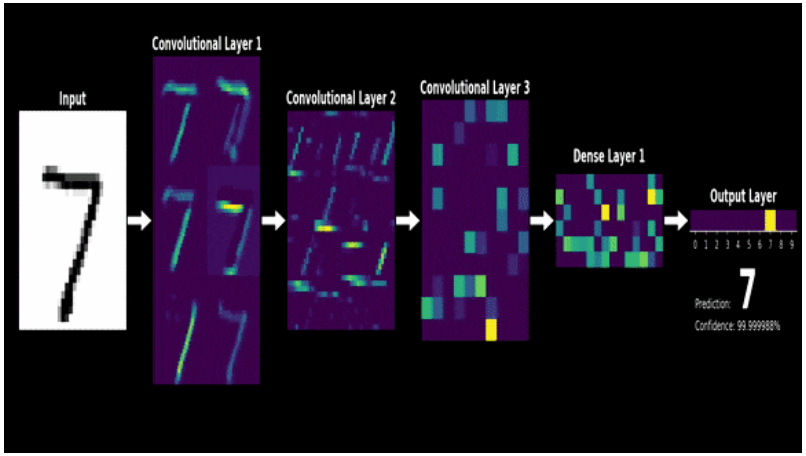
研究工作一：模型可视化

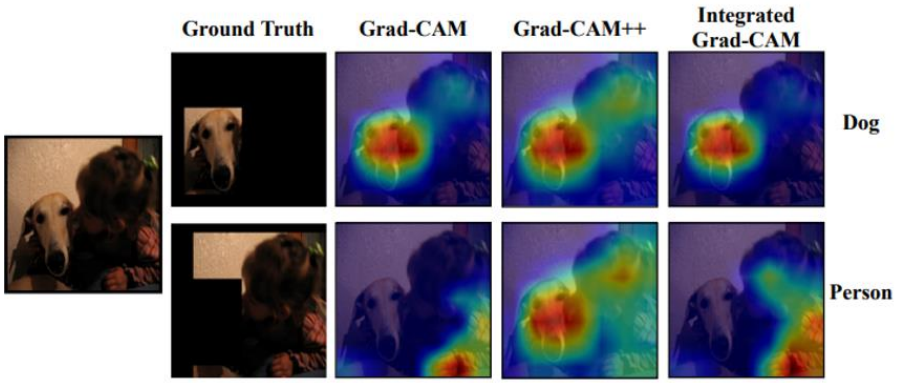

开源、鲁棒、可信的说话人识别
开源
鲁棒
可信
多场景
采集流程
多场景
挑战问题
跨场景
泛化场景
复杂场景
模型
评测
多场景中文明星
数据库
跨、泛化、复杂场景
说话人识别
模型可视化与
性能评测

研究工作一：模型可视化

计算机视觉中的模型可视化

- 诸多可视化工具用来辅助解释模型判决的可信度。
- 通过生成显著图 (saliency maps) 来查明一幅图片中的哪些区域对模型判决起到了关键性作用。
- 人可以容易地理解一幅图片的显著图，从而可以评判可视化工具的特质。

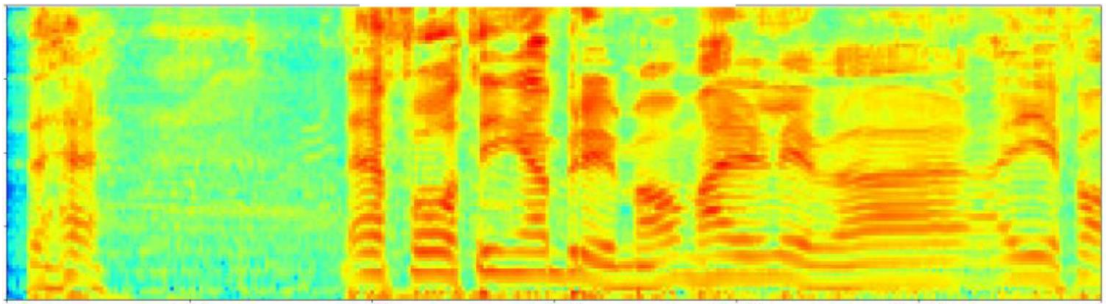




从视觉到听觉

说话人识别中的模型可视化

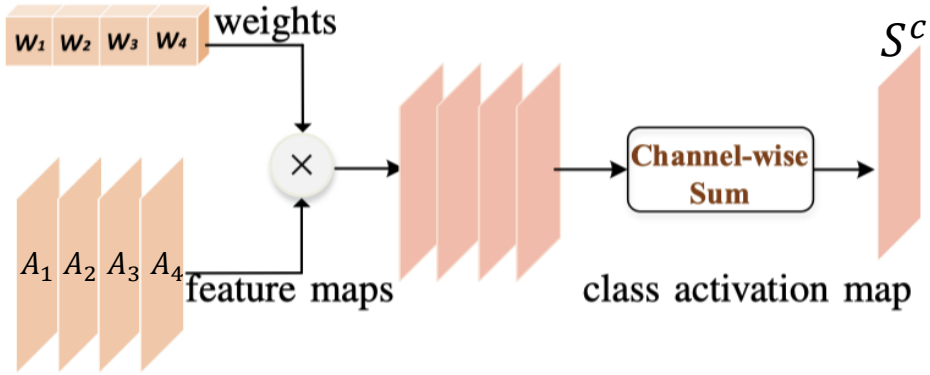
- 不同于图像，语音是可“听”不可“看”，导致人们对语音显著图的理解和评判变得极为困难，相关研究工作尚少。
- 仅有的研究工作大都是拿来主义，并不确信所用可视化工具对于说话人识别任务是否可靠，使得从可视化中所得到的结论不能完全令人信服。



研究目标：针对说话人识别任务，寻找一种可靠的模型可视化工具

三种类激活图可视化方法

类激活图 (CAM)



$$S^c = \mathrm{ReLU}(\sum_k w_k^c \cdot A_k)$$

Grad-CAM

1). 目标类别后验概率

$$y^c = f_c(x;\theta)$$

2). 第k个激活图权重

$$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Score-CAM

1). 目标类别的激活图

$$A^k = f_c(x;\theta)$$

2). 第k个激活图权重

$$\hat{x}_k = x \circ \{\mathrm{Norm}(\mathrm{Upsampling}(A^k))\}$$

$$w_k^c = f(\hat{x}_k)$$

Layer-CAM

1). 目标类别后验概率

$$y^c = f_c(x;\theta)$$

2). 第k个激活图权重

$$w_{ij}^{kc} = relu(\frac{\partial y^c}{\partial A_{ij}^k})$$

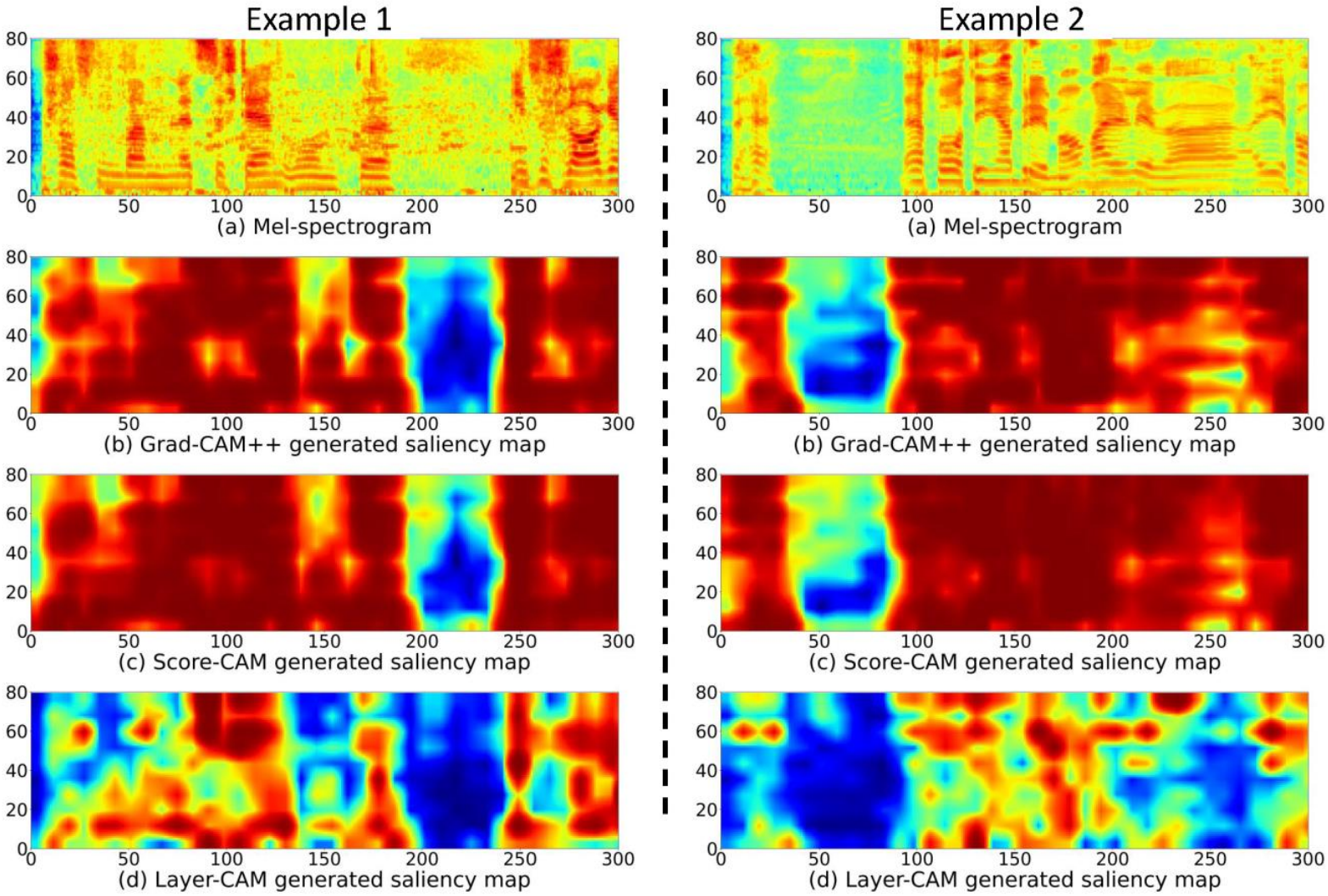
先进的深度说话人表征模型

- ResNet34SE

网络层	模块	输出
输入	–	80×200×1
Conv2D	3×3×32, Stride 1	80×200×32
ResNetBlock1	$\begin{bmatrix} 3\times3\times32 \\ 3\times3\times32 \\ \text{SE Layer} \end{bmatrix}$ ×3, Stride 1	80×200×32
ResNetBlock2	$\begin{bmatrix} 3\times3\times64 \\ 3\times3\times64 \\ \text{SE Layer} \end{bmatrix}$ ×4, Stride 2	40×100×64
ResNetBlock3	$\begin{bmatrix} 3\times3\times128 \\ 3\times3\times128 \\ \text{SE Layer} \end{bmatrix}$ ×6, Stride 2	20×50×128
ResNetBlock4	$\begin{bmatrix} 3\times3\times256 \\ 3\times3\times256 \\ \text{SE Layer} \end{bmatrix}$ ×3, Stride 2	10×25×256
Pooling	TSP[174]	20×256
Flatten	–	5120
Dense	–	256
Dense	AM-Softmax[138]	5994

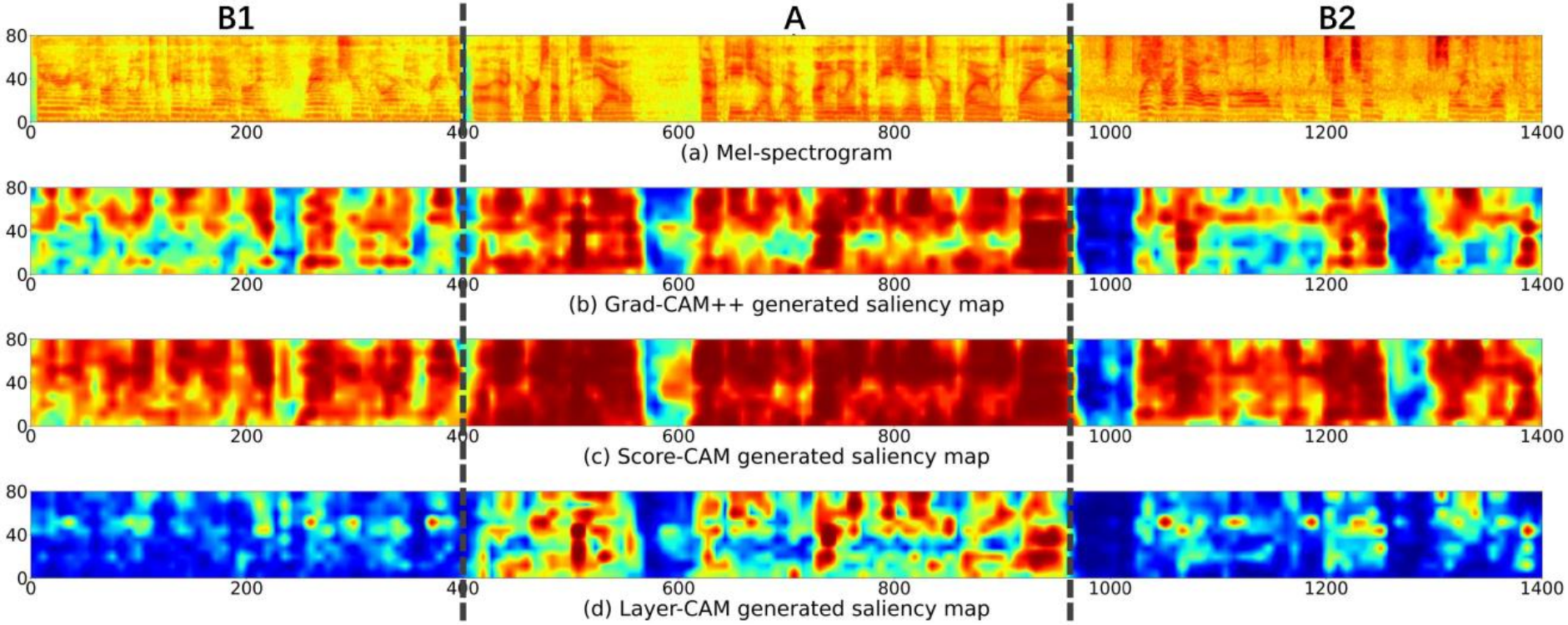
单个说话人的类激活图

- 语音段和非语音片段可分；没有明确的时频模式



多个说话人的类激活图

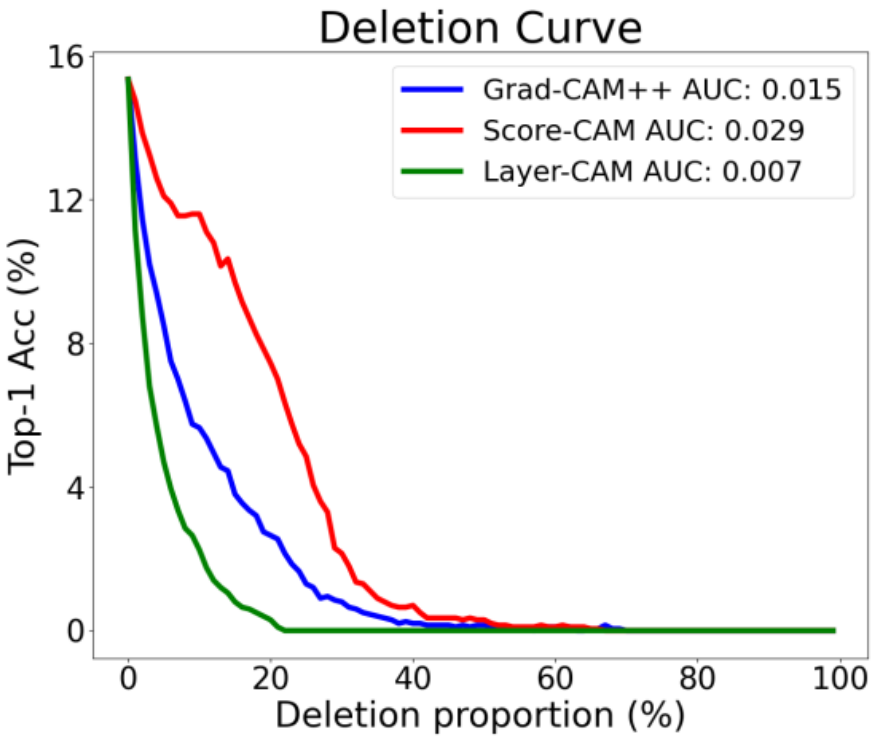
- A是目标说话人，B是干扰说话人

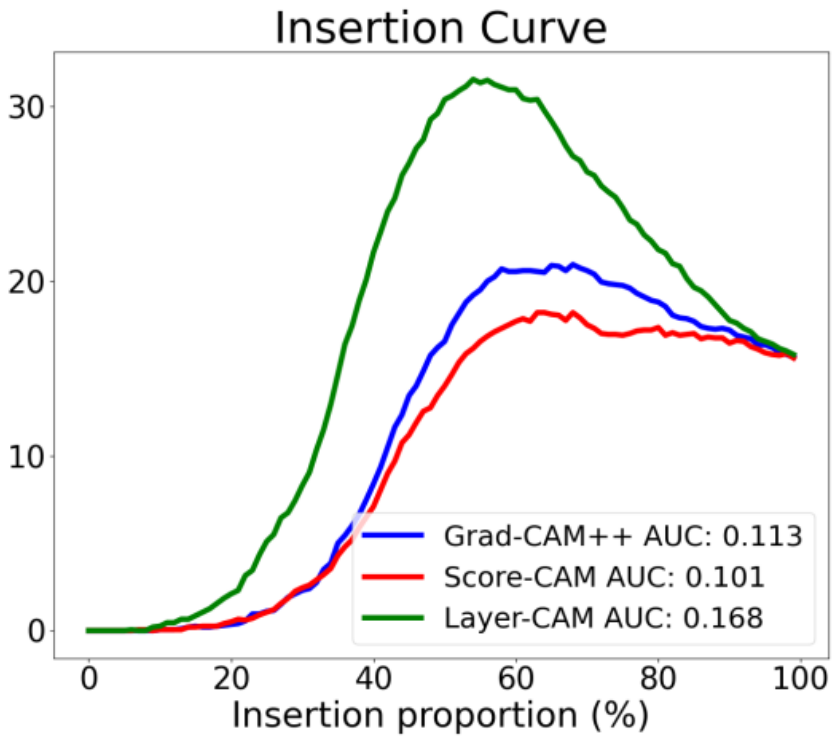


- 直观上看，Layer-CAM 效果最佳，可以准确地定位目标说话人的语音片段，并且几乎完美地掩盖了非目标说话人的语音片段

删除和插入实验

B-A-B 测试





- 对比 AUC，Layer-CAM 效果最佳。

目标说话人定位与识别

- 第一步，定位目标说话人所在的位置
  第二步，仅使用定位后的语音片段进行目标说话人识别

Cases	A-B			A-B-A			B-A-B			A-B-C		
Original	49.15%			83.55%			15.35%			30.30%		
Settings	G-CAM	S-CAM	L-CAM	G-CAM	S-CAM	L-CAM	G-CAM	S-CAM	L-CAM	G-CAM	S-CAM	L-CAM
S1	43.00%	34.00%	6.75%	75.55%	62.50%	8.90%	15.15%	12.10%	4.15%	22.40%	17.45%	3.90%
S2	46.60%	46.60%	61.85%	79.90%	79.25%	85.00%	15.70%	16.00%	35.20%	26.85%	26.85%	45.15%
S3	48.45%	48.60%	49.40%	82.65%	82.35%	80.15%	15.75%	16.00%	20.20%	29.20%	29.55%	31.05%
S4	49.20%	48.25%	53.20%	82.10%	82.65%	82.90%	17.20%	16.15%	24.05%	30.10%	29.20%	34.65%
S4+S3	48.65%	48.15%	51.15%	82.50%	82.25%	82.60%	16.50%	16.15%	21.90%	29.40%	29.30%	33.55%
S4+S3+S2	48.55%	48.40%	59.85%	82.20%	82.00%	87.30%	16.10%	16.20%	28.65%	29.65%	29.15%	42.75%
S4+S3+S2+S1	47.70%	47.50%	**71.55%**	81.50%	80.65%	**92.20%**	16.10%	16.10%	**44.60%**	27.95%	27.45%	**58.90%**

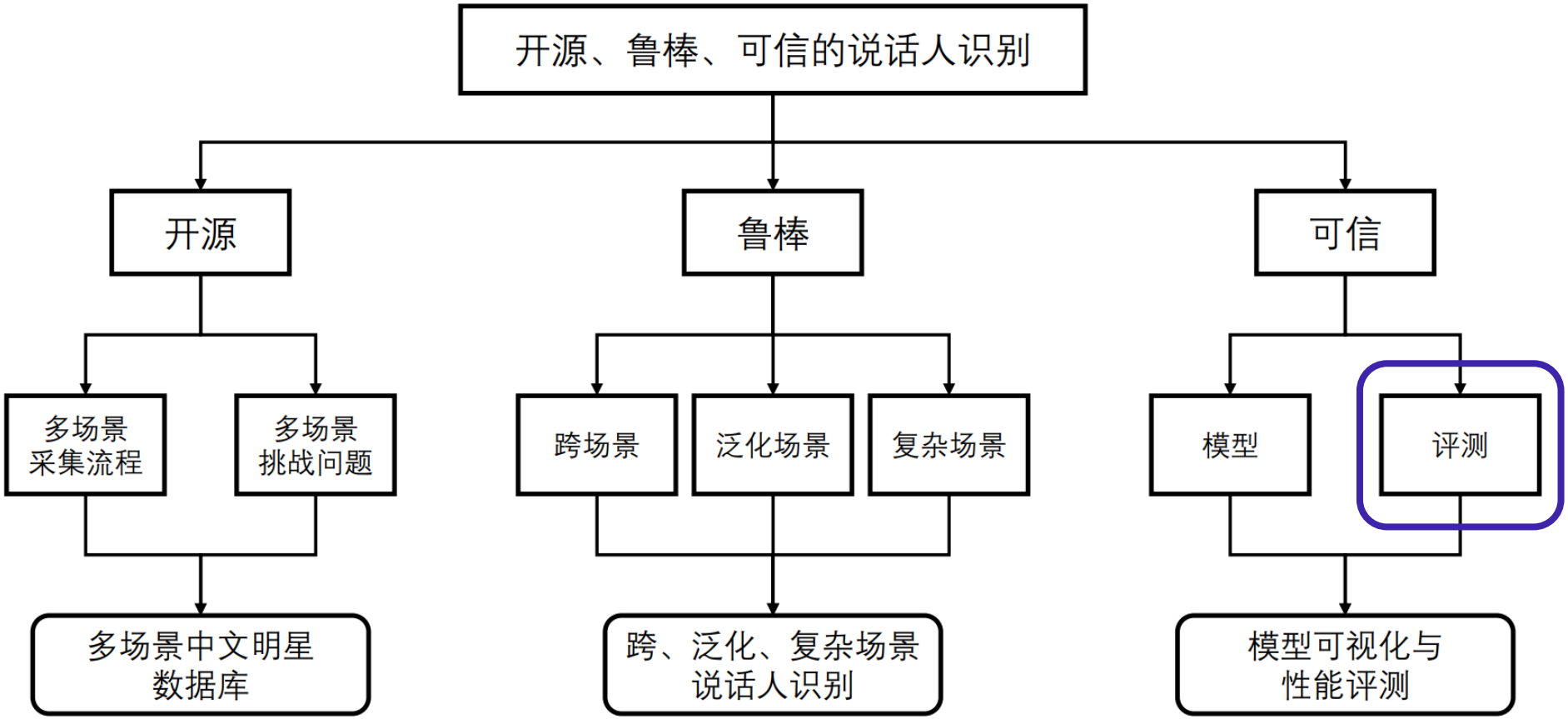
- Layer-CAM 取得了明显的性能提升，表明 Layer-CAM 可以识别重要的说话人区分性区域，而其他两种 CAM 则不能。
- Layer-CAM 聚合来自不同卷积层的显著图可以进一步提高识别性能。这一现象与特征聚合技术一致，而其他两种 CAM 没有这种趋势。
- 综上，Layer-CAM 是三种 CAM 方法中唯一可靠的可视化工具。

小结

模型可视化

- 通过定性和定量分析，确定了 Layer-CAM 是一个可靠的可视化工具，可用来理解、解释深度说话人模型。

研究工作二：性能评测


开源、鲁棒、可信的说话人识别
开源
鲁棒
可信
多场景
采集流程
多场景
挑战问题
跨场景
泛化场景
复杂场景
模型
评测
多场景中文明星
数据库
跨、泛化、复杂场景
说话人识别
模型可视化与
性能评测

二、性能评测

基准评测 vs. 用户体验

- 基准评测 VoxSRC 2021 第一名

System Index	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H		VoxSRC20-dev		VoxSRC21-val	
	EER(%)	$DCF_{0.01}$	EER(%)	$DCF_{0.01}$	EER(%)	$DCF_{0.01}$	EER(%)	$DCF_{0.05}$	EER(%)	$DCF_{0.05}$
S1	0.5249	0.0498	0.7130	0.0627	1.1240	0.0923	1.8330	0.0867	1.6020	0.0906
S2	0.5037	0.0356	0.6435	0.0514	**0.9737**	0.0783	1.5760	0.0753	**1.3350**	0.0685
S3	0.4613	**0.0232**	**0.6342**	**0.0477**	0.9932	0.0763	**1.4770**	0.0726	1.4550	0.0813
S4	0.5673	0.0309	0.6759	0.0550	1.0360	0.0830	1.5860	0.0797	1.4620	0.0776
S5	**0.4401**	0.0253	0.6518	0.0494	0.9914	0.0738	1.4960	**0.0691**	1.3610	**0.0628**
S6	0.4825	0.0374	0.6707	0.0508	1.0270	0.0783	1.5160	0.0725	1.4050	0.0730
S7	0.4825	0.0283	0.6511	0.0484	0.9965	0.0738	1.4910	0.0699	1.4180	0.0660
S8	0.5090	0.0340	0.6587	0.0489	0.9954	**0.0707**	1.4940	0.0699	1.4180	0.0698
S9	0.5673	0.0461	0.6961	0.0584	1.0910	0.0856	1.7040	0.0845	1.6420	0.0942
Fusion										
S1~S9	**0.4189**	**0.0217**	**0.5826**	**0.0414**	**0.8868**	**0.0630**	**1.3400**	**0.0624**	**1.2710**	**0.0590**

基准评测与用户体验之间的性能差距显著

两类研究思路

数据复杂性

- 基本假设：性能差距归因于声学失配问题。
- 解决方案：精心设计了各种基准数据集来模拟现实生活中的各种声学场景；例如，HI-MIA、NIST SRE、VoxCeleb 和 CN-Celeb。

测试列表

- 测试列表用于探测和度量说话人识别系统的性能。
- 如果测试列表设计不当，则无法恰当地度量目标系统的性能。
- 是本研究的关注点。

基准测试列表 vs. 实际测试列表

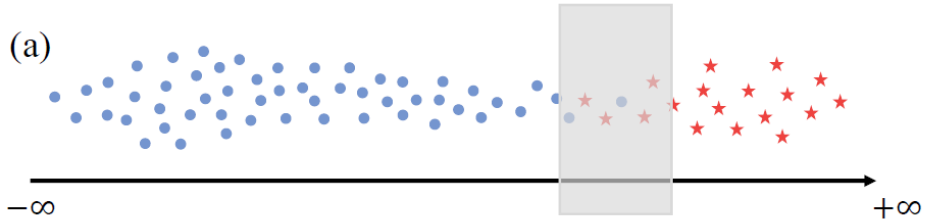
基准测试列表

- 全交叉配对，闯入测试数远多于目标测试数
- 存在大量的简单测试，尤其是闯入测试

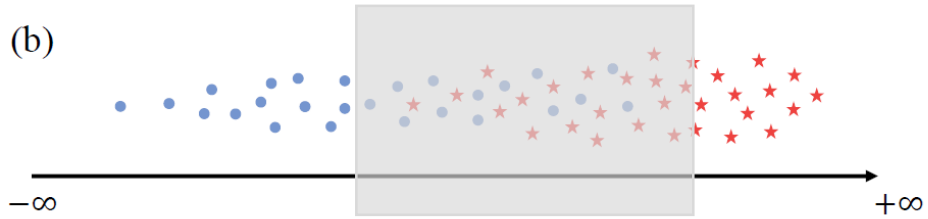
实际测试列表

- 通常目标测试数多于闯入测试数
- 闯入测试更具挑战性

基准评测
分数分布



实际测试
分数分布



二者之间存在测试偏差

定义两个新概念

测试配置

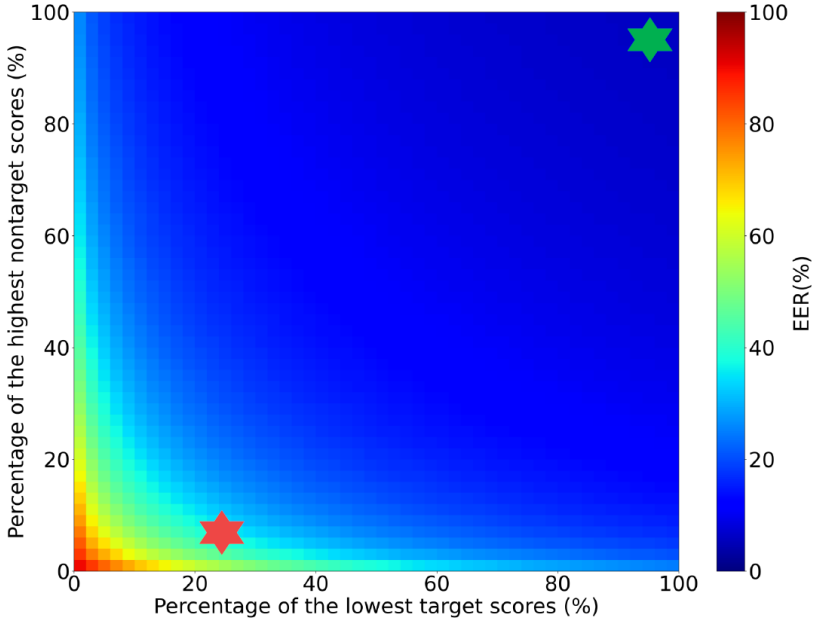
- 给定一组注册/测试语音，测试配置定义为一个测试列表的子集，用于目标系统性能评测。
- 全交叉配对是最大的测试配置，测试列表涵盖了所有的测试样例。
- 不同测试配置的性能不同，反映了目标系统在不同部署情况下的性能。

配置-性能分布图

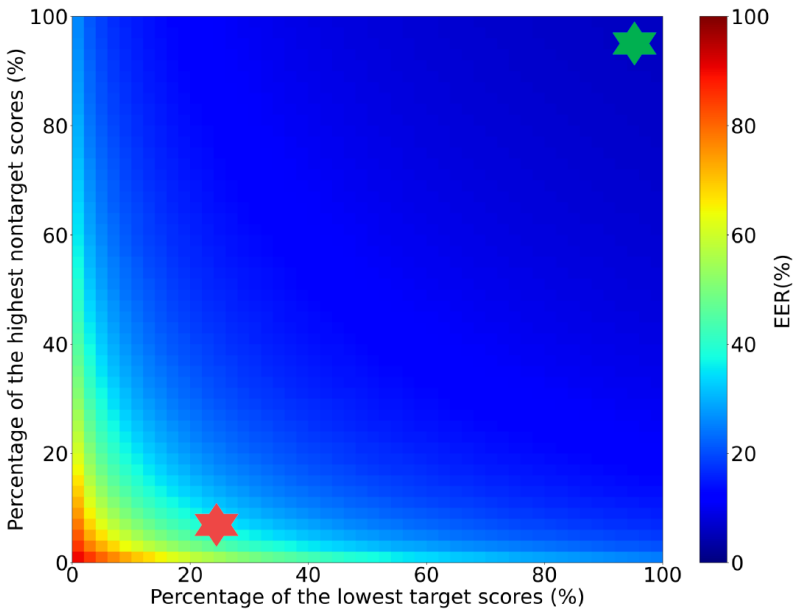
- 收集所有测试配置及其性能，便可更全面地评价目标系统的综合能力。
- 通过一个配置-性能分布图来呈现。其中，x 轴对应于目标测试样例的子集，y 轴对应于闯入测试样例的子集。
- 图中每个位置 (x, y) 对应于一个特定的测试配置，其颜色深浅代表性能好坏。

配置-性能分布图

- 合理选择设计测试配置，使分布图具有空间结构属性
  - 对于某个目标系统，根据其判决分数对测试样例进行排序；然后从有序列表中选择测试样例来构建测试配置。
  - 对于目标测试列表 (x 轴)，从左到右逐渐选择得分较高的测试样例；对于闯入测试列表 (y 轴)，从下到上逐渐选择分数较低的测试用例。
  - 显然，左下区域的测试配置比右上区域的测试配置更难。



观察现象



- 大面积的高性能区域揭露了基准测试列表中存在大量的简单测试样例。
- 两组测试配置 (红星和绿星) 分别对应于实际测试和基线测试。
- 两组测试配置有着截然不同的性能表现，解释了基准评测与用户体验性能差距的原因。

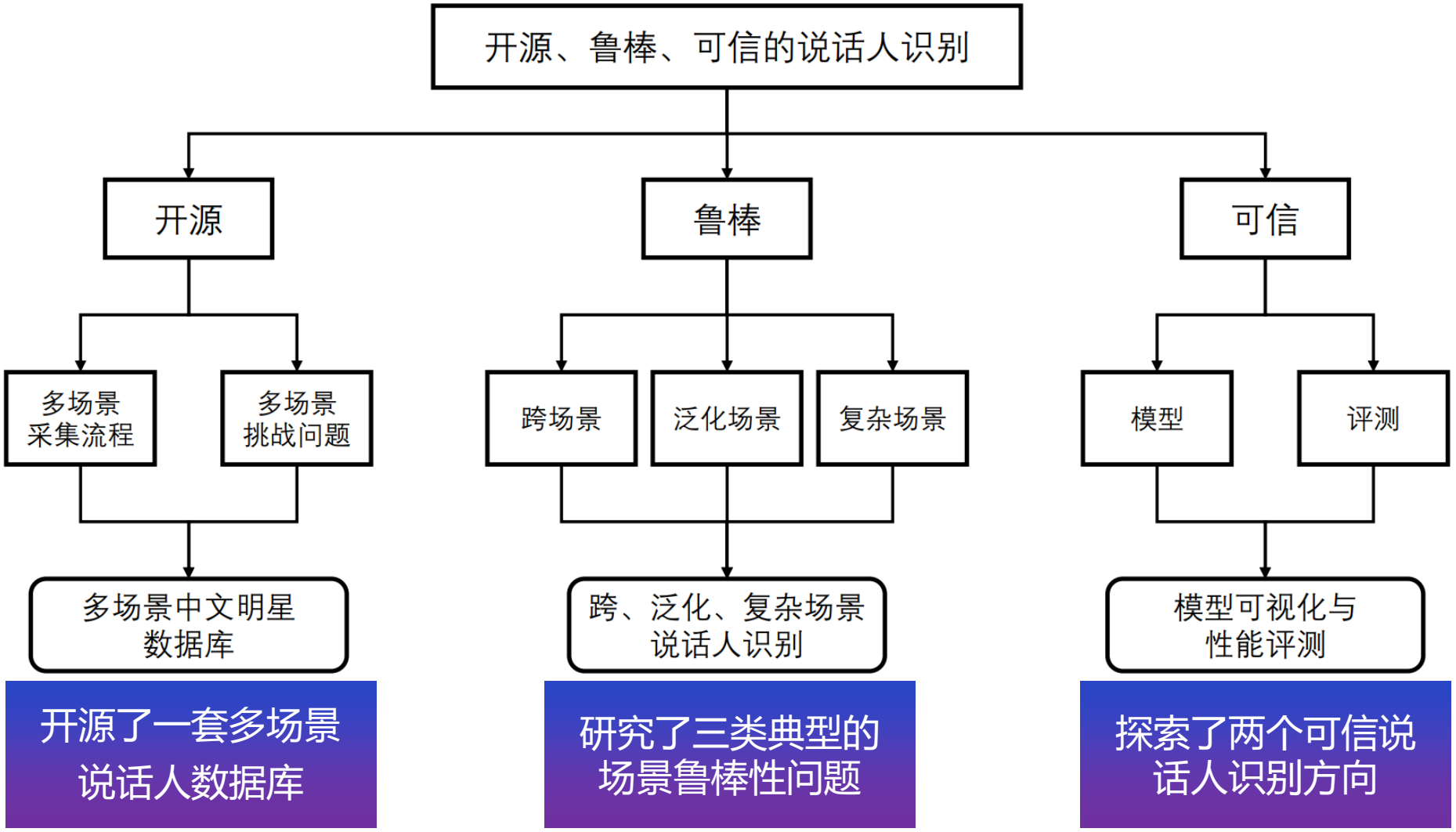
小结

可信性能评测

- 从基准评测与用户体验不一致的现象出发，分析了测试列表偏差问题。
- 定义了测试配置，设计了配置-性能分布图，解释了基准评测与实际测试不符的原因。

总结与展望

工作总结


开源、鲁棒、可信的说话人识别
开源
鲁棒
可信
多场景
采集流程
多场景
挑战问题
跨场景
泛化场景
复杂场景
模型
评测
多场景中文明星
数据库
跨、泛化、复杂场景
说话人识别
模型可视化与
性能评测
开源了一套多场景
说话人数据库
研究了三类典型的
场景鲁棒性问题
探索了两个可信说
话人识别方向

总结与展望

未来展望

- **开源：**首先，开源更智能化、自动化的多场景数据采集平台，包括搜索下载、一键处理、在线标注等功能模块；其次，开源上万级的多场景说话人数据库，提供更丰富的场景类型、更规范的标注协议。

- **鲁棒：**基于开源多场景数据，组织 CNSRC 2022 说话人识别竞赛，开辟新的研究方向和评测任务。例如，多复杂场景下的说话人识别、大规模目标说话人检索等。

- **可信：**探索更多可靠的可视化工具，理解深度模型的内在机制，反馈设计更合理的模型结构。分析配置-性能图中不同位置的模式表现，更全面的分析系统真实性能，挖掘系统的潜在风险。

代表性研究成果

项目支持

- 第64批中国博士后科学基金面上资助 一等奖
- 国家自然科学基金重点项目 (61633013)
- 清华大学-浦发银行数字金融技术联合研究中心


中国博士后科学基金资助证书
Certificate of China Postdoctoral Science Foundation Grant

清华大学 博士后研究人员 李蓝天

（全国博管办编号为 210895 ），获得第 64 批中国博士后科学基金面上资助 – 等资助，资助编号为 2018M640133 。

特颁此证。

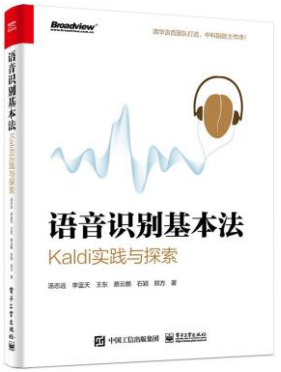
中国博士后科学基金会

2018 年 11 月 20 日

证书查验请登录中国博士后科学基金会网站

论文专利

- 合著学术专著 1部《语音识别基本法》
- 发表学术论文 19篇；SCI 3篇、EI 16篇。
- 申请发明专利 8项、软著 1项



荣誉获奖

- 清华大学博士后 b2 支持计划, 2018.07 – 2020.07
- 中国电子学会科学技术奖 (技术发明类一等奖), 第五完成人, 2022.01

感谢合作导师郑方教授！

感谢各位评审老师！

李 蓝 天

清华大学 语音和语言技术中心

http://cslt.riit.tsinghua.edu.cn/