面向会议场景的声纹识别技术

张鹏远

中国科学院声学研究所

中科院语言声学与内容理解重点实验室

2020年11月21日

提纲

一 研究背景及挑战

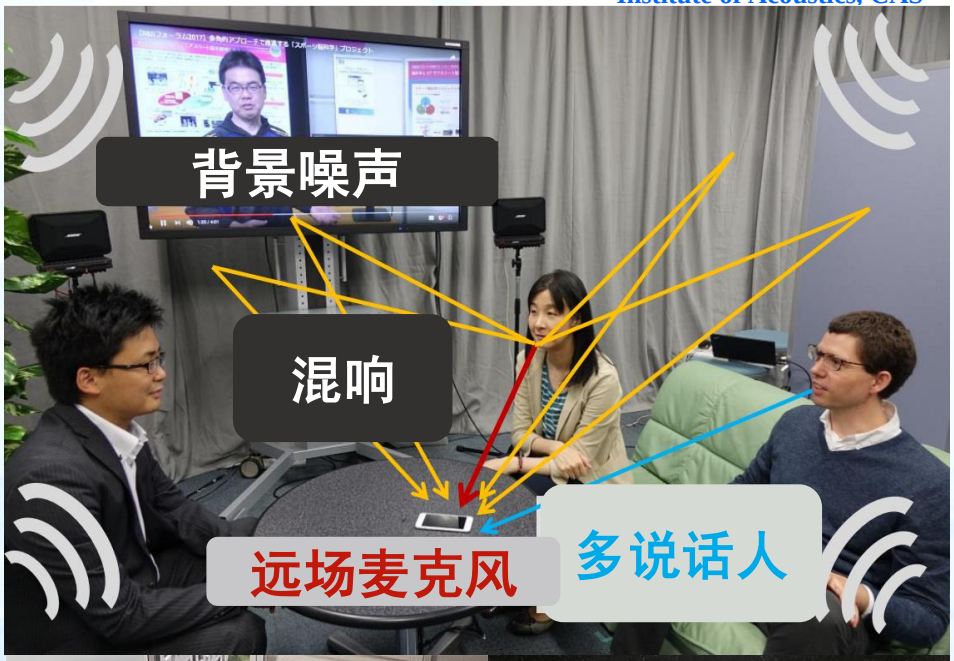
二 多人会话场景下的说话人聚类

三 多人会话场景下的说话人分离

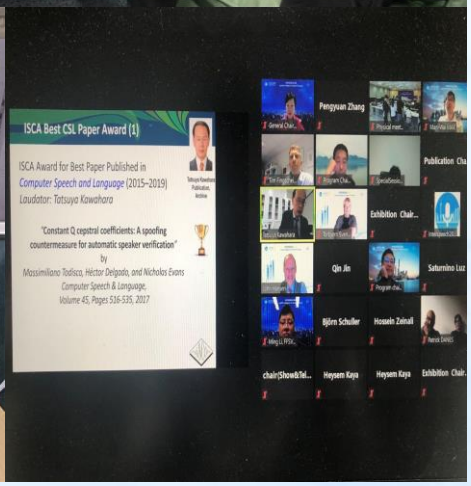
四 跨域声纹识别

研究背景及挑战

- 背景
  - 多人会议身份自动识别
  - 海量数据类别和内容的精确管理



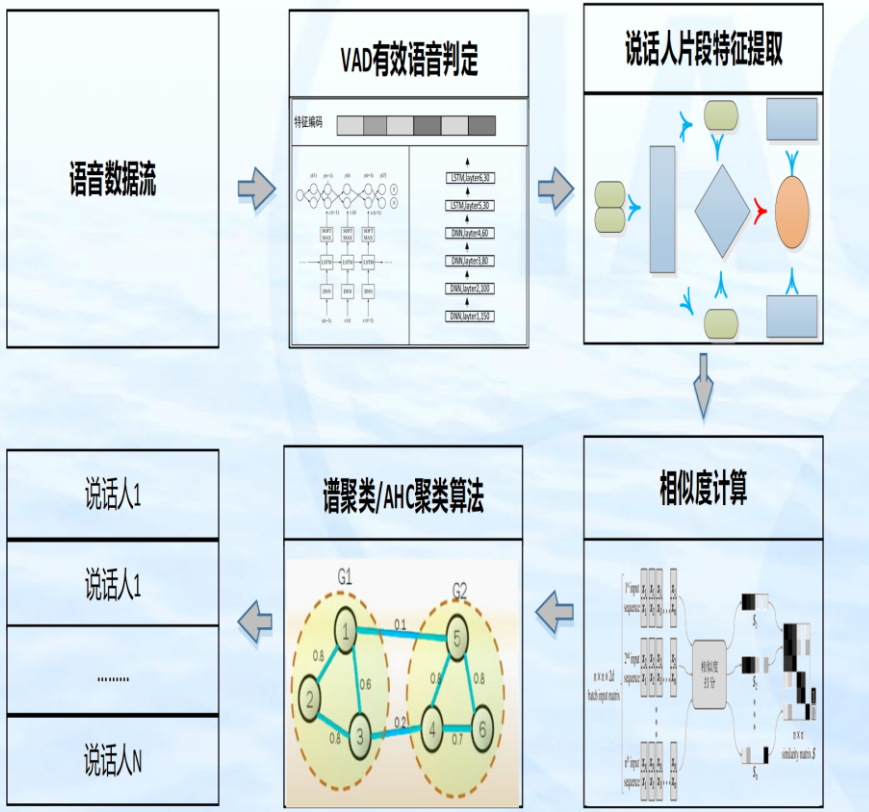
- 挑战
  - 领域差异普遍存在
    - 信道的差异（电话互联网）
    - 设备的差异（麦克风种类）
    - 环境的差异（干净嘈杂、近场远场）
    - 语言的差异（中英、方言）
  - 多数情况下，目标领域缺少有标注的训练数据



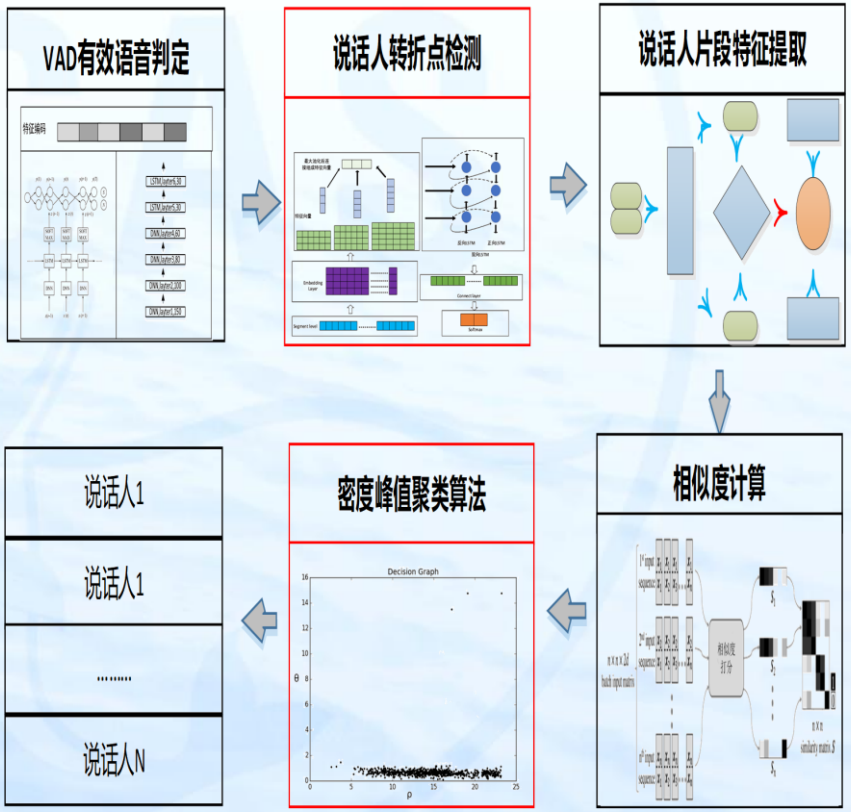
多人会话场景下的说话人聚类：框架

中国科学院声学研究所
Institute of Acoustics, CAS

传统多人会话分离数据流程



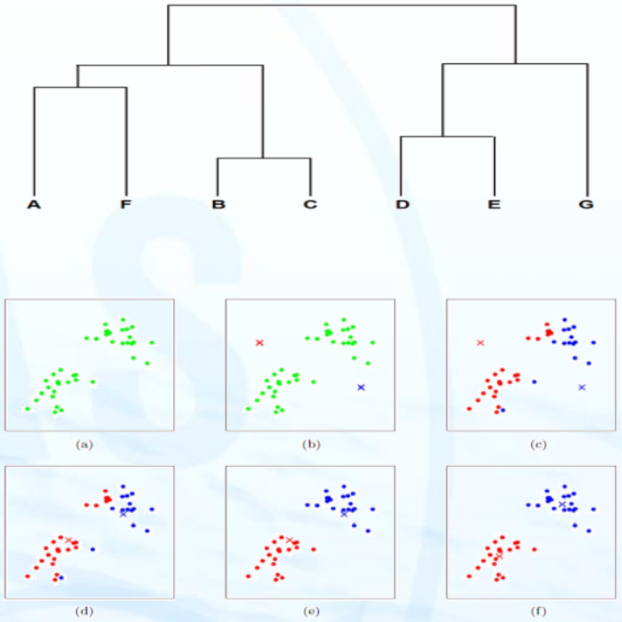
改进多人会话分离数据流程

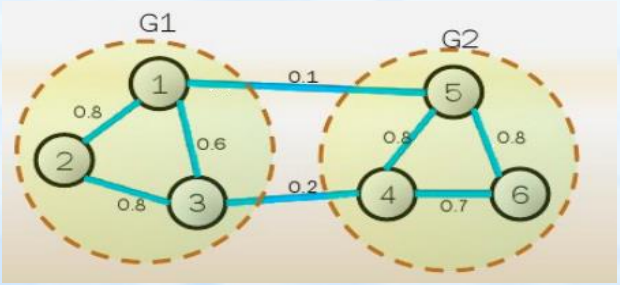


多人会话场景下的说话人聚类：传统方法

中国科学院声学研究所
Institute of Acoustics, CAS

- 传统聚类方法
  - 已知类数聚类方法
    - 层次聚类方法、K-means聚类方法
    - 优点：性能好、实现简单
    - 缺点：运行速度较慢、需预先设计类别数谱聚类方法
  - 未知类数聚类方法
    - 谱聚类
    - 优点：避免高维向量造成的奇异性问题、易于实现
    - 缺点：不适用于语料不平衡的数据集





多人会话场景下的说话人聚类：密度峰值聚类

密度峰值聚类算法

- 算法思想
  - 类簇中心点的密度大于周围邻居点的密度
  - 类簇中心点与更高密度点之间的距离相对较大
- 参数设定
  - 局部密度ρ_i
  - 与高密度之间的距离 δ

Rodriguez, Alex, and Alessandro Laio. "Clustering by fast search and find of density peaks." *Science* 344.6191 (2014): 1492-1496.

多人会话场景下的说话人聚类：密度峰值聚类

中国科学院声学研究所
Institute of Acoustics, CAS

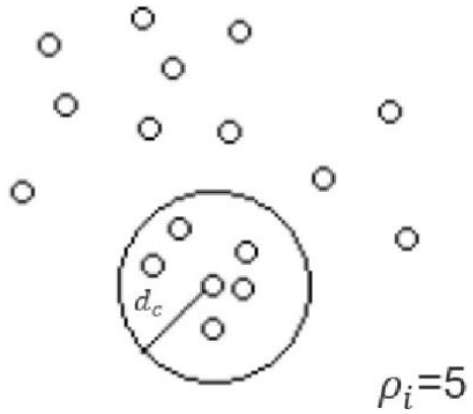
密度峰值聚类算法

参数计算

- 局部密度$\rho_i = \sum_j \chi(d_{ij} - d_c)$，$\chi(k) = \begin{cases} 1 & k \le 0 \\ 0 & k > 0 \end{cases}$

  ρ_i瞎徵仰踿棚焛i盦踿棚山仰dc盦焛盦二旒

  d_c 通常取值为所有d_{ij}升序排列中第2%个d_{ij}的值



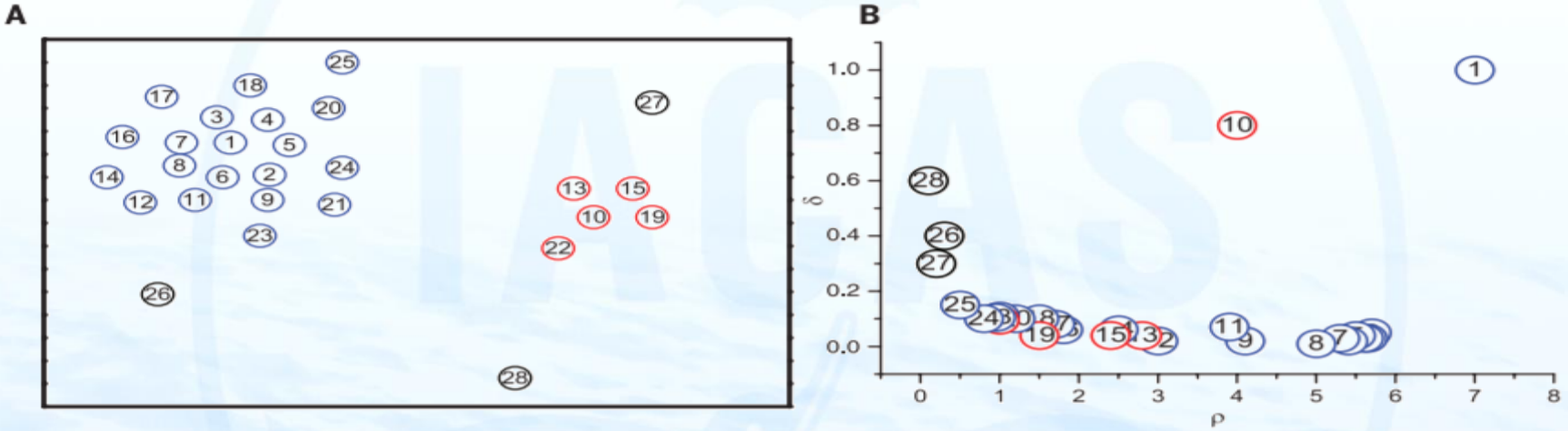
- 聚类中心距离 δ
  - ✓ 将每个点的密度从大到小依次排列
  - ✓ 先确定密度最大的点（i点）的聚类中心距离，聚类中心距离δ_i为与i点最远的点（n点）距离d_{in}
  - ✓ 确定其他点的聚类中心距离：其他点的聚类中心距离是等于在密度大于该点的点集合中，与该点距离最小的的那个距离
  - ✓ 依次确定所有的聚类中心距离δ

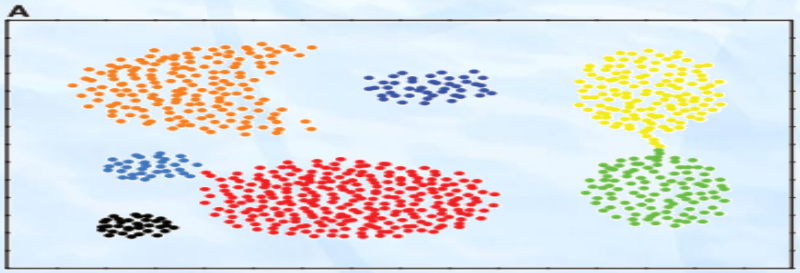
$$\begin{cases} \delta_i = \min\limits_{\rho_j > \rho_i}(d_{ij}) & , i \ge 1 \\ \delta_i = \max\limits_{j}(d_{ij}) & , \rho_i \text{为全局最高} \end{cases}$$

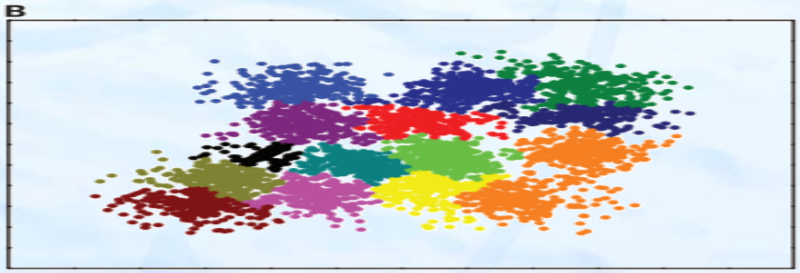
多人会话场景下的说话人聚类:密度峰值聚类

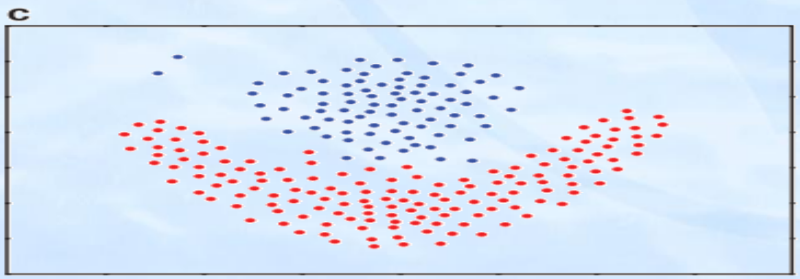
中国科学院声学研究所
Institute of Acoustics, CAS

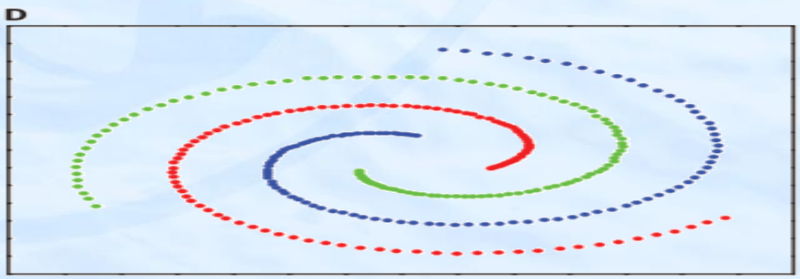
❑ 密度峰值聚类算法











多人会话场景下的说话人聚类:密度峰值聚类

中国科学院声学研究所
Institute of Acoustics, CAS

密度峰值聚类算法

- 密度峰值聚类挂接相似矩阵算法流程
  - Step1：计算各个语音片段之间相似度→相似矩阵S[n][n]
  - Step2：遍历相似矩阵每一行：S[i][j] = S[i][i] - S[i][j]
  - Step3：依次求出各个点的局部密度$\boldsymbol{\rho}$与密度距离$\boldsymbol{\delta}$
  - Step4：针对所有i点 $\mathrm{r}[i] = \rho_i * \delta_i$
  - Step5：针对 **r** 从大到小排序
  - Step5：针对i从1到 $n_{\max\_spk}$： $k[i] = r[i+1]/r[i]$
  - Step6：$n_{spk} = Arg(Max(k[i]))$
  - Step7：针对所有数据点完成类别归属

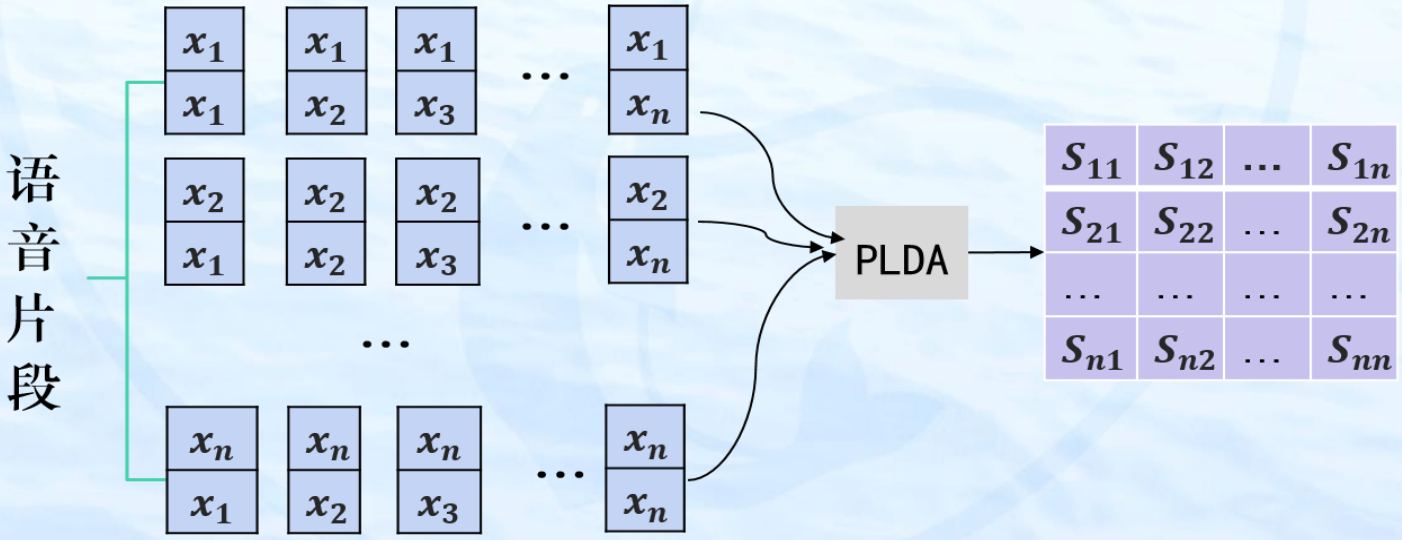
多人会话场景下的说话人聚类:密度峰值聚类

中国科学院声学研究所
Institute of Acoustics, CAS

密度峰值聚类算法

- 语音片段相似度矩阵S[n][n]计算
  - 相似矩阵S[i][j]代表音频片段i与音频片段j之间的相似度

$$S_i = [S_{i1}, S_{i2}, \dots S_{in}] = f_{PLDA}(\begin{bmatrix} x_i \\ x_1 \end{bmatrix}, \begin{bmatrix} x_i \\ x_2 \end{bmatrix}, \dots \begin{bmatrix} x_i \\ x_n \end{bmatrix})$$



多人会话场景下的说话人聚类：密度峰值聚类

实验数据库

- NIST2010电话数据集

  NIST电话数据：8355个音频文件，总大小为37.7G,总时长为702小时

- 阿波罗数据集

  Train数据：128个音频文件，时常30分钟，人数：4-61人

  Dev数据： 30个音频文件，时常30分钟，人数：7-61人

- 选取动机
  - 验证DPCA算法在稳定两人电话数据上的聚类性能
  - 验证DPCA算法在未知人数且低信噪比(0-20db)数据上的聚类性能

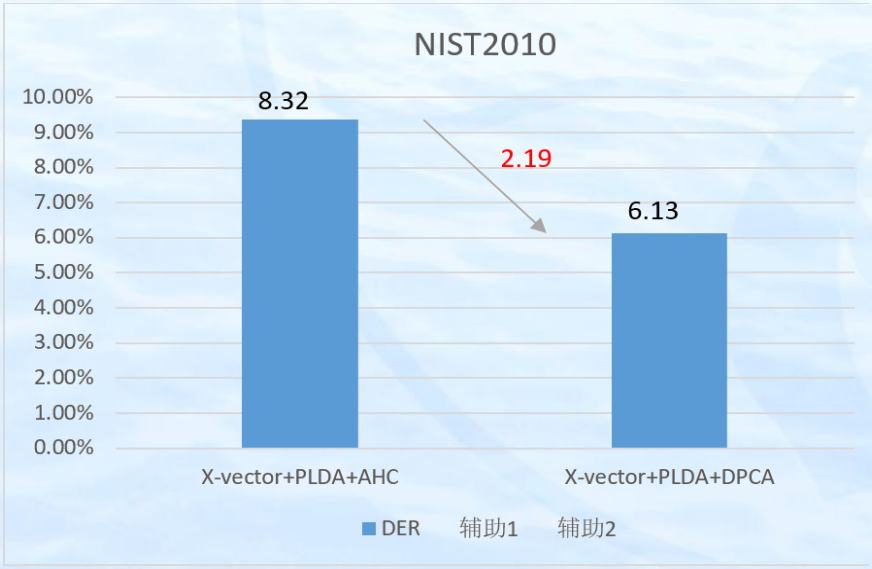
多人会话场景下的说话人聚类：密度峰值聚类

中国科学院声学研究所
Institute of Acoustics, CAS

实验结果与分析

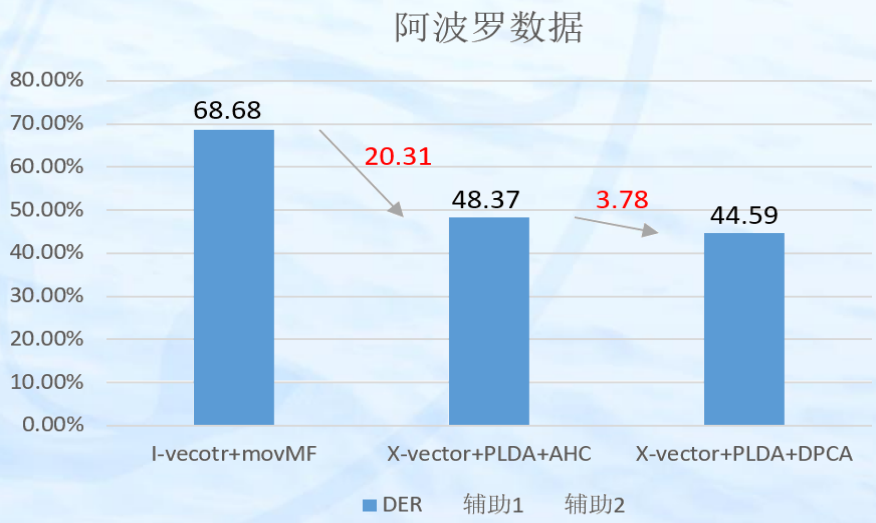
- 在NIST2010-tel数据上DER绝对下降2.19个点
- 在阿波罗数据集上DER绝对下降3.78个点

一些小的音频片段被正确分类



未知人数时计算的人数更贴近实际

语料不平衡时更鲁棒



多人会话场景下的说话人聚类：结果分析

中国科学院声学研究所
Institute of Acoustics, CAS

实验结果与分析

算法优点

- 能快速检测出数据集中的类数
- 善于处理不规则形状的簇，对语料不均衡的数据集效果更好
- 善于发现球状簇，对参数选择不敏感

存在问题

- 密度峰值聚类依赖于相似矩阵的构建
- 硬切分(1.5s窗长、0.75s窗移)片段不纯导致相似度矩阵计算不精确

改进方法

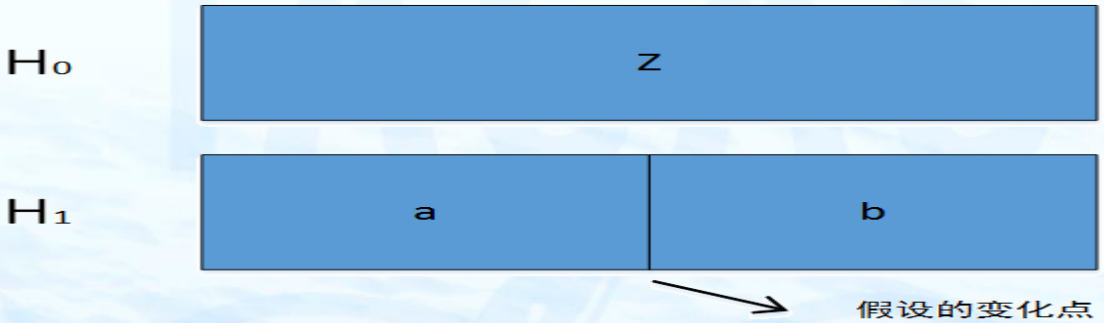
- 增加说话人转折点检测模块提高语音片段纯度

多人会话场景下的说话人聚类：转折点检测

中国科学院声学研究所
Institute of Acoustics, CAS

传统基于距离度量准则

- BIC贝叶斯信息距离
- GLR广义似然比距离



传统转折点检测优缺点

- 无监督、实现简单、不需要说话人的先验信息
- 门限需提前在开发集划定
- 产生的片段比较碎短，不利于后续聚类
- 语音背景、信道差异影响大不鲁棒

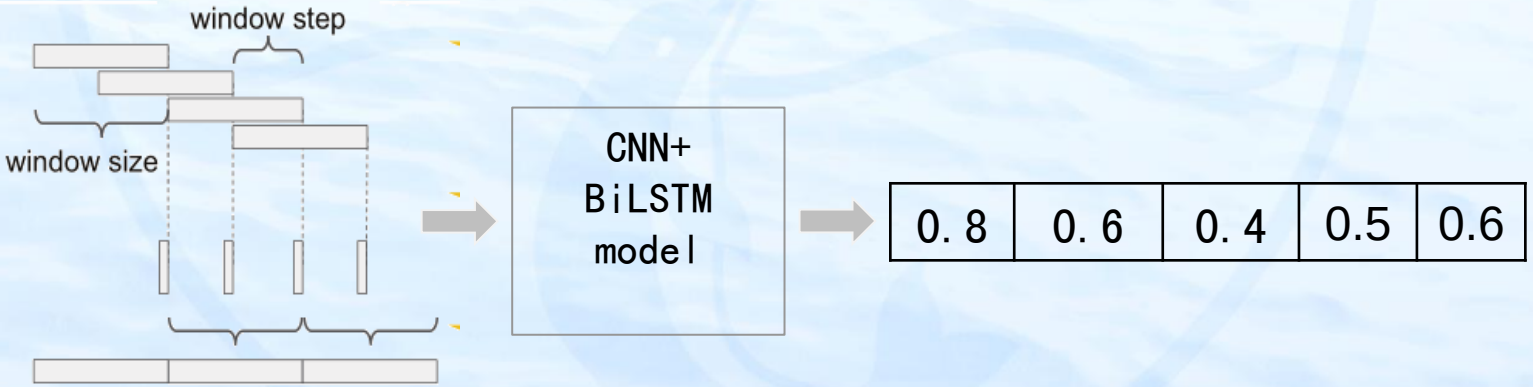
多人会话场景下的说话人聚类：转折点检测

中国科学院声学研究所
Institute of Acoustics, CAS

基于CNN+Bi-LSTM卷积的说话人转折点检测

- 针对过完VAD的语音片段，按照0.25秒的长度切分成小的语音片段，判断相邻片段之间的语音相似度，根据语音相似度的大小来判定是否存在说话人说话人转折点

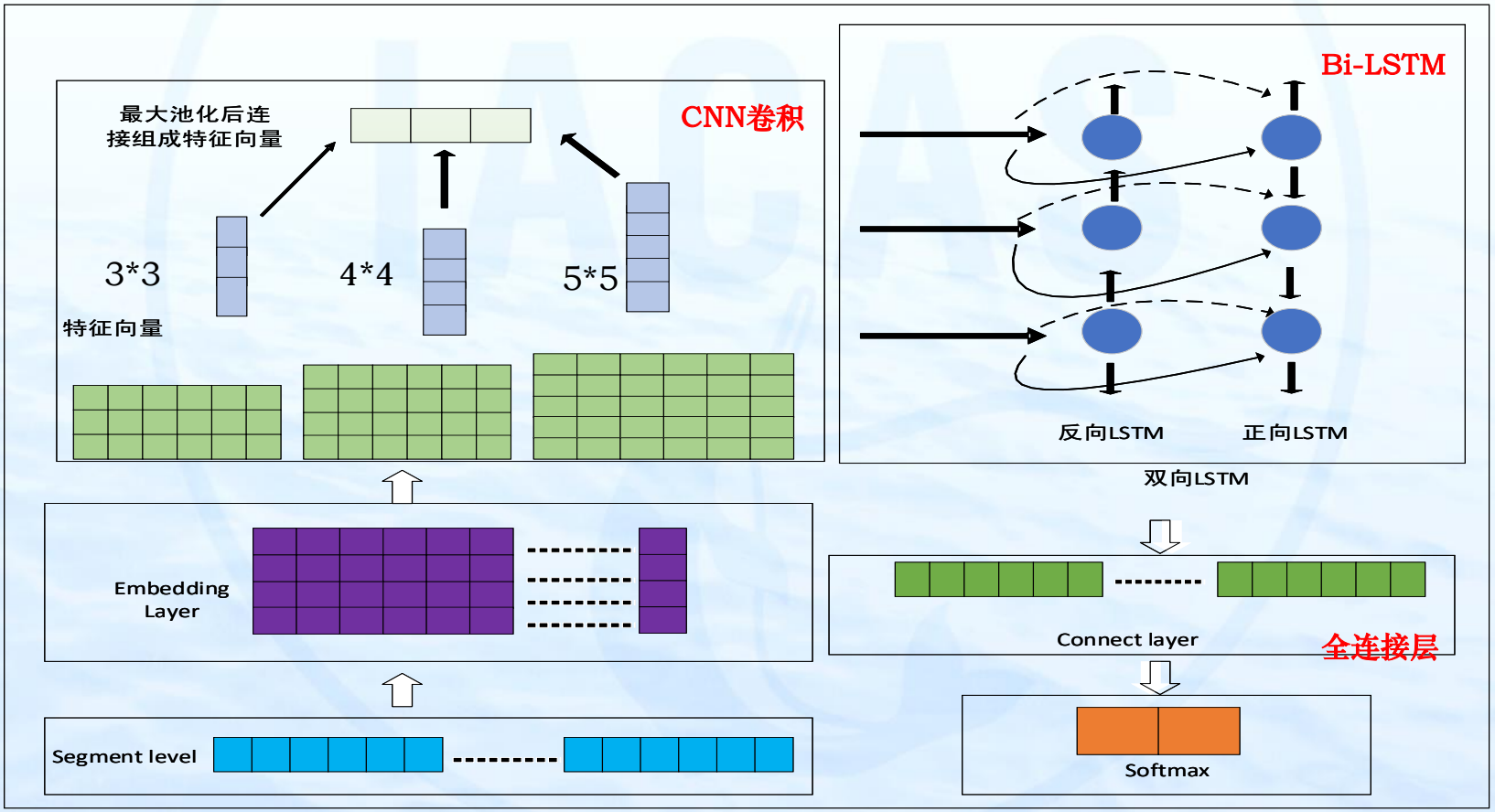
- 语音相似度得分处于低谷且小于设定阈值



多人会话场景下的说话人聚类：转折点检测

中国科学院声学研究所
Institute of Acoustics, CAS

说话人转折点检测网络



多人会话场景下的说话人聚类：转折点检测

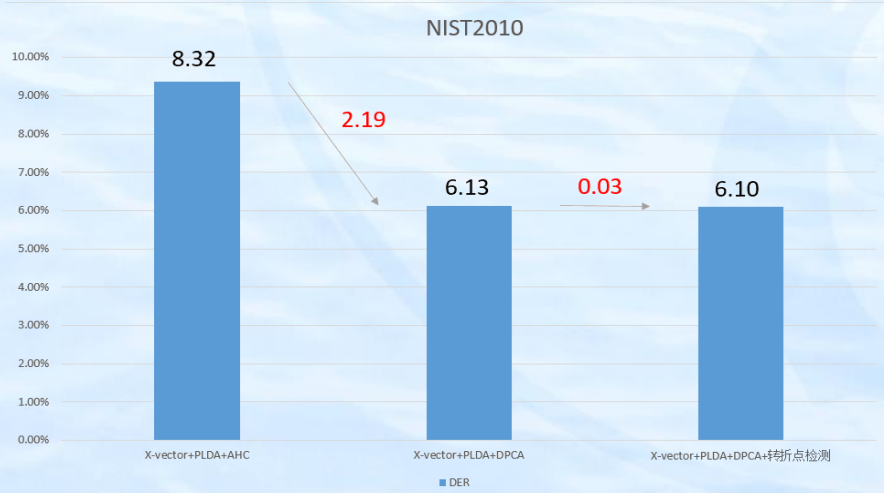
中国科学院声学研究所
Institute of Acoustics, CAS

实验结果与分析

- 在NIST2010-tel数据上DER绝对下降0.03个点
- 在阿波罗数据集上DER绝对下降10.06个点

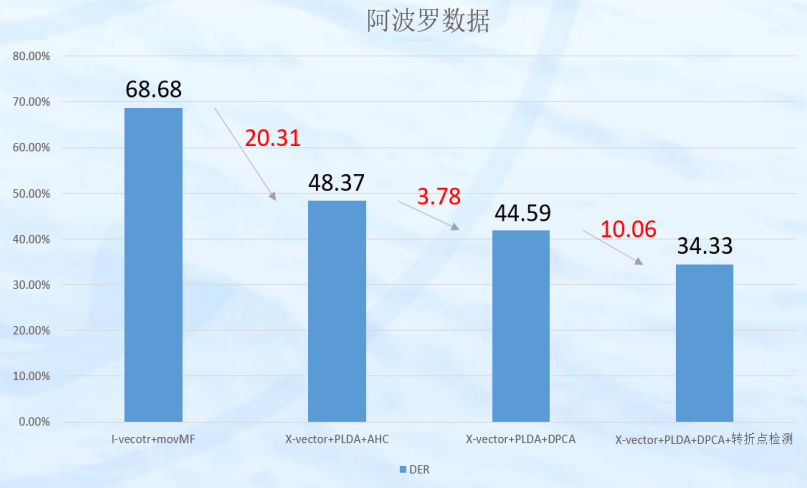
NIST2010电话数据比较干净
自带的标注相对比较精确

针对干净数据，说话人转折点
检测并未带来负面影响



阿波罗数据人数单条4-61人
信噪比比较低VAD无法有效切分

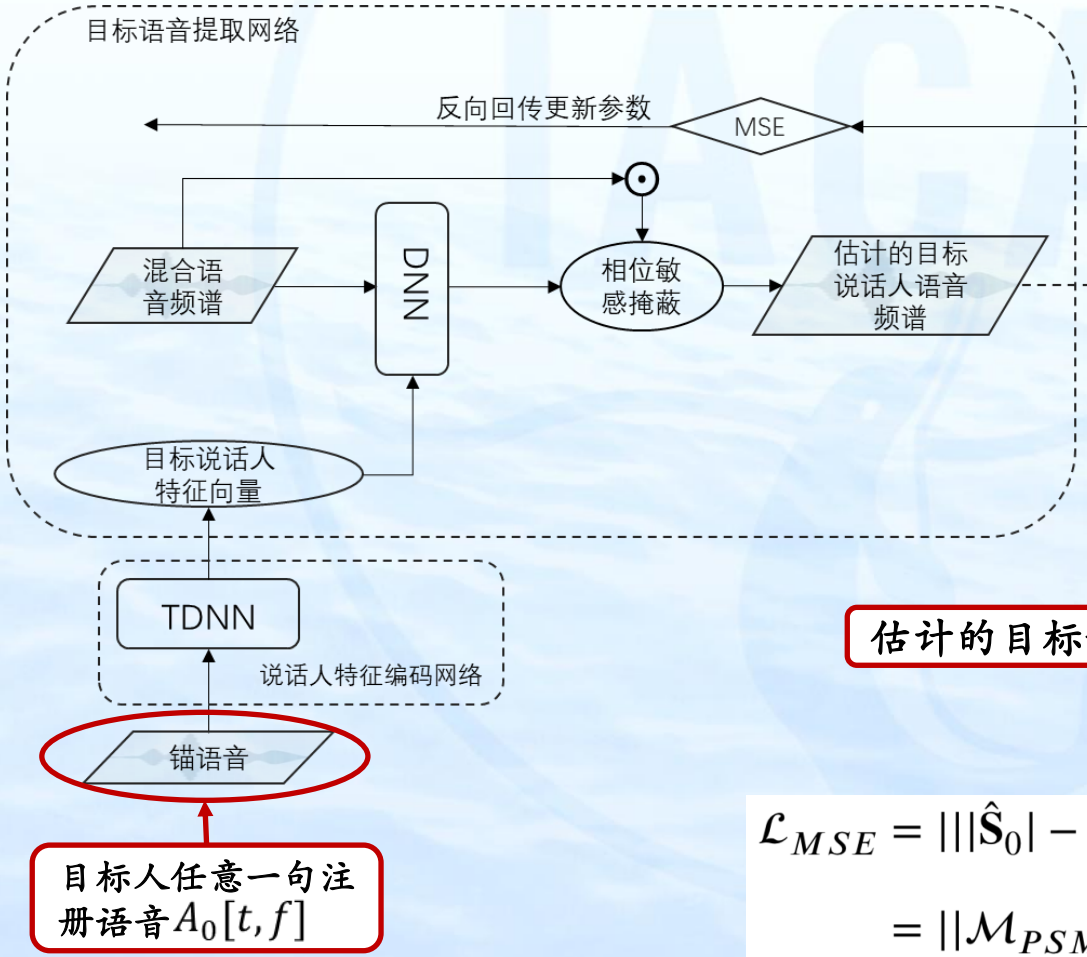
针对多人说话人的有效语音片
段都可以相对精确的分割出来



多人会话场景下的说话人分离：目标人语音提取

中国科学院声学研究所
Institute of Acoustics, CAS

目标语音提取模型结构 (target speech extraction network , TEnet)



- 说话人特征能够编码不同人之间的差异，将其作为目标人先验信息。
- 从锚语音中获取。

信号模型

目标语音　干扰语音　噪声

$$Y[t,f] = S_0[t,f] + \sum_{i=1}^{I-1} S_i[t,f] + Z[t,f]$$

估计的目标语音　相位敏感掩蔽

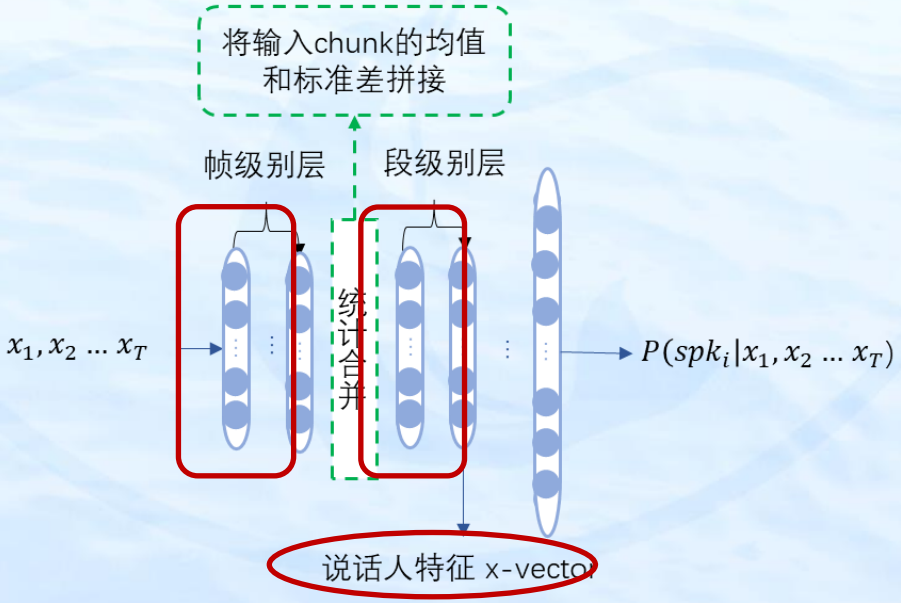
$$\hat{\mathbf{S}}_0 = \mathcal{M}_{PSM} \odot \mathbf{Y}$$

$$\mathcal{L}_{MSE} = |||\hat{\mathbf{S}}_0| - |\mathbf{S}_0| \odot max(0, \cos(\theta_{\mathbf{y}} - \theta_{\mathbf{s}_0}))||^2$$

$$= ||\mathcal{M}_{PSM} \odot |\mathbf{Y}| - |\mathbf{S}_0| \odot max(0, \cos(\theta_{\mathbf{y}} - \theta_{\mathbf{s}_0}))||^2$$

多人会话场景下的说话人分离：目标人语音提取

- x-vector 说话人特征编码网络
  - 先训练一个说话人分类网络，再从网络隐层提取说话人特征
  - 网络结构中增加了一个数据统计合并层，用来计算当前输入段（chunk）的统计量
  - 将模型段级别层的激活值作为说话人特征向量x-vector

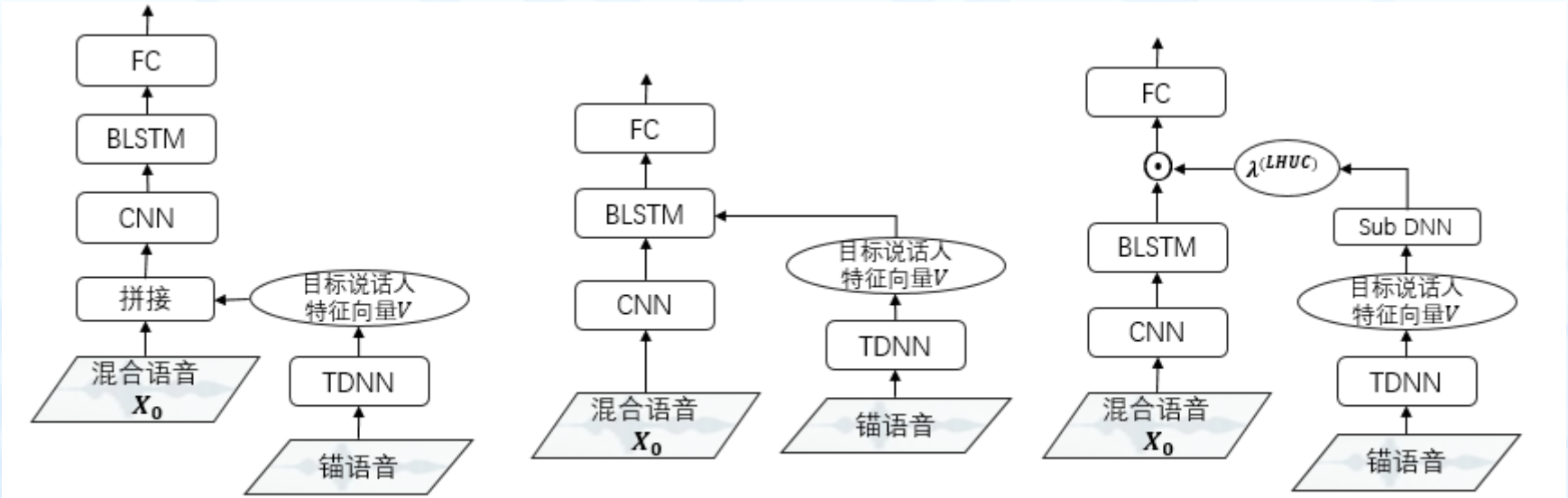


多人会话场景下的说话人分离：目标人语音提取

中国科学院声学研究所
Institute of Acoustics, CAS

说话人特征的引入方式

- TEnet 的模型主体部分采用CNN+BLSTM+FC 的结构
- 探究了三种不同的引入目标说话人特征*V*的方式：输入补偿，多层级引入，自适应隐层贡献调节的方式



$$X_{CNN} = \sigma_0(L_{CNN}([X_0, V]; \psi_{CNN}))$$

$$X_{BLSTM} = \sigma_k(L_{BLSTM}(X_{CNN}; \psi_{BLSTM}))$$

$$X_{CNN} = \sigma_0(L_{CNN}(X_0; \psi_{CNN}))$$

$$X_{BLSTM} = \sigma_k(L_{BLSTM}([X_{CNN}, V]; \psi_{BLSTM}))$$

$$V^{(LHUC)} = \sigma_{sub}(L_{sub}(V; \psi_{sub}))$$

$$X_{BLSTM}^{LHUC} = \sigma^{LHUC}(V^{(LHUC)} \odot X_{BLSTM})$$

多人会话场景下的说话人分离：目标人语音提取

实验设置

数据集

- 说话人编码网络：SWBD 和 SRE数据集， 7k个不同说话人。
- TEnet：
  - WSJ0-2mix：两个说话人0-5dB混合语音。
  - WSJ0-2mix-noise：在WSJ0-2mix数据基础上增加MUSAN噪声（音乐，风声，脚步声，动物声，雨声等）。

说话人编码网络

- 7层TDNN网络：5层帧级别层-统计合并层-2层段级别层，512维。
- $\mathrm{TDNN1}^{\mathrm{frame}}_{\{-2,-1,0,1,2\}}\mathrm{TDNN2}^{\mathrm{frame}}_{\{-2,0,2\}}\mathrm{TDNN3}^{\mathrm{frame}}_{\{-3,0,3\}}\mathrm{TDNN4}^{\mathrm{frame}}_{\{0\}}\mathrm{TDNN5}^{\mathrm{frame}}_{\{0\}} -$
  $-\ \mathrm{statistic}(1500-3000)--\mathrm{TDNN6}^{\mathrm{chunk}}_{\{0\}}\mathrm{TDNN7}^{\mathrm{chunk}}_{\{0\}}$

TEnet

- CNN：2*卷积网络[64*卷积核[9*9]]。
- BLSTM：每个方向3*640维-循环节点128维。
- FC：1层，输入1280维，输出257维。

评估指标

- SDR：信号失真比，语音质量。
- WER：词错误率，语音识别结果（声学模型 5*650 TDNN）。

多人会话场景下的说话人分离：目标人语音提取

中国科学院声学研究所
Institute of Acoustics, CAS

❑ 引入方式对比

说话人特征引入方式	SDR	WER
-	1.12	82.35
输入补偿	9.84	42.96
多层级输入	**10.17**	**41.93**
自适应隐层贡献调节	9.93	43.40

- TEnet框架能够有效提取目标语音，大幅提升目标语音的SDR。
- 多层级输入性能较好：先通过CNN 提取频谱的高层表征，再将它与x-vector 拼接之后输入给BLSTM 模型进行语音提取。

• 锚语音时长对比

锚语音	SDR	WER
-	1.12	82.35
short	9.78	44.64
middle	9.82	42.71
long	**10.17**	**41.93**

6.1%↓

short — 锚语音的时长小于3 秒
middle — 锚语音的时长在3-10 秒之间
long — 锚语音的时长大于10 秒

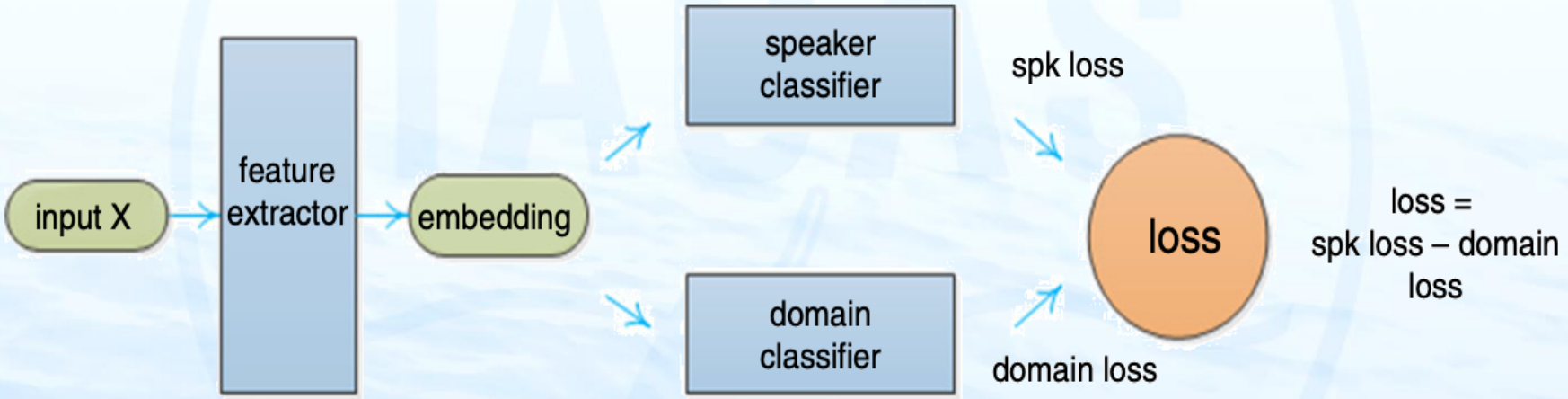
- 锚语音的时长对目标语音提取的结果有较大影响，总的来看，时长越大，TEnet提取出语音的质量越好。
- 利用长的锚语音进行语音提取能够获得相对于短的锚语音6.1% 的WER 下降。

跨域声纹识别：有监督自适应（DAT）

中国科学院声学研究所
Institute of Acoustics, CAS

基于对抗学习的有监督领域自适应

- 领域对抗训练（Domain Adversarial Training，DAT）

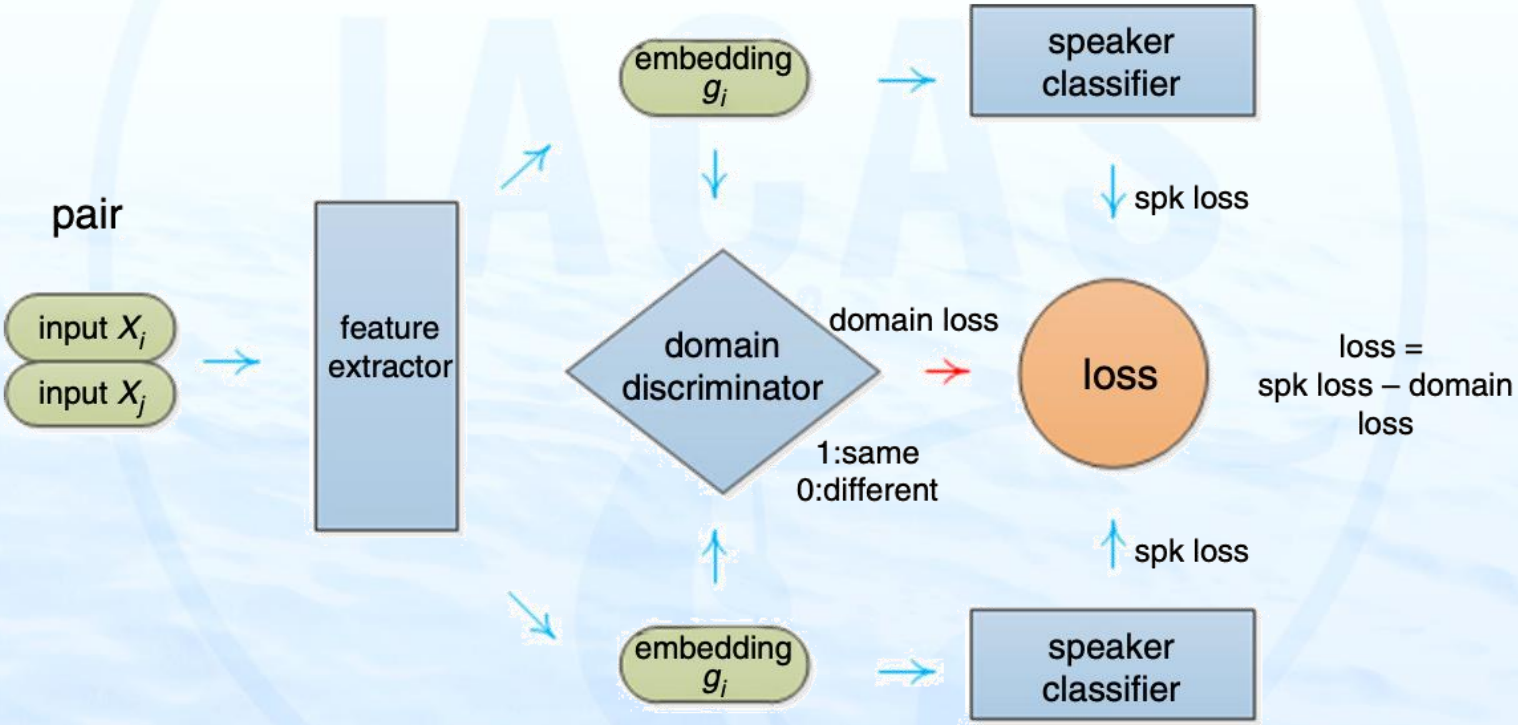


- 基于multi-task框架，说话人分类器与领域分类器相互对抗，驱使网络更关注语音中的说话人信息、削弱领域信息

跨域声纹识别：有监督自适应（DAS）

中国科学院声学研究所
Institute of Acoustics, CAS

- 基于对抗学习的有监督领域自适应
  - 孪生对抗网络（Domain Adversarial Siamese network，DAS）



  - 成对输入；领域分类器修改为判决器，提取更纯粹的领域无关特征

跨域声纹识别：DAS 实验结果

中国科学院声学研究所
Institute of Acoustics, CAS

- 基于对抗学习的有监督领域自适应
  - 已开展的实验
    - 远场实验（AISHELL-wakeup）
      - 实验1: 训练 1m+5m，评估 1m注册--5m测试

No	System	Loss	EER(%)
1	Base	\	5.78
2	DAT	spk	3.72
3		spk - domain	3.39
4	DAS	spk	3.72
5		spk - domain	3.37

1m-5m训练数据量1:1
DAT 领域二分类
DAS 领域异同0-1判决
原理等价，性能几乎相同
Domain loss贡献约10%性能提升

跨域声纹识别：DAS 实验结果

中国科学院声学研究所
Institute of Acoustics, CAS

- 基于对抗学习的有监督领域自适应
  - 已开展的实验
    - 远场实验（AISHELL-wakeup）
      - 实验2: 训练 0m+5m，评估 0m注册--5m测试

No	System	Loss	EER(%)
1	Base	\	8.92
2	DAT	spk	5.28
3		spk - domain	5.31
4	DAS	spk	5.28
5		spk - domain	4.71

0m-5m训练数据量1:16
DAT 领域分类器训练失衡，domain loss没有作用
DAS domain loss稳定贡献10%性能提升

跨域声纹识别：DAS 实验结果

中国科学院声学研究所
Institute of Acoustics, CAS

- 基于对抗学习的有监督领域自适应
  - 已开展的实验
    - 远场实验（AISHELL-wakeup）
      - 实验3: 训练 0m+1m+5m，评估 0m注册--3m测试

No	System	Loss	EER(%)
1	Base	\	9.69
2	DAT	spk	5.01
3		spk - domain	5.17
4	DAS	spk	5.01
5		spk - domain	4.44

测试集出现未知领域
DAS相比于DAT，更能提取到“领域无关”的语音特征

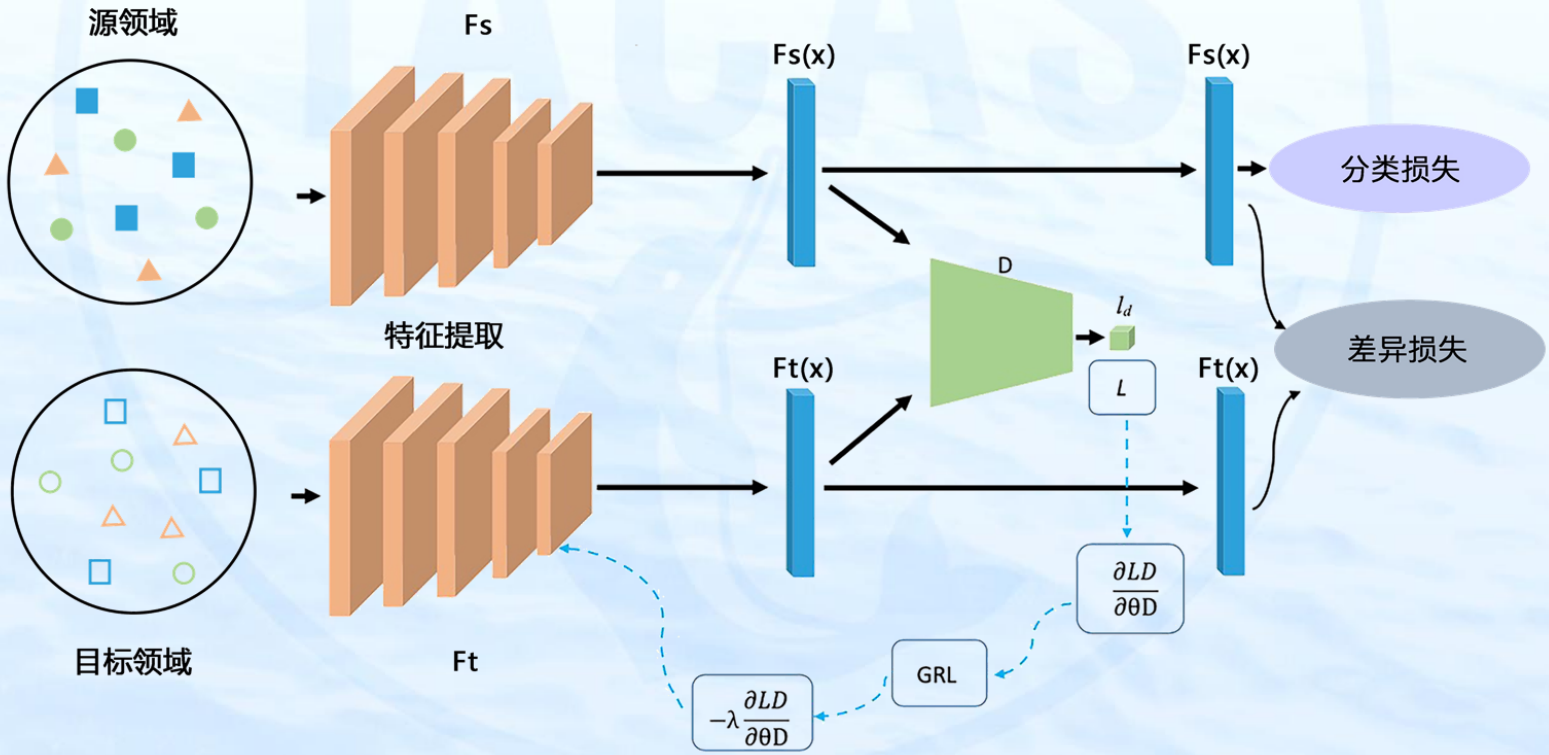
跨域声纹识别： 无监督自适应

中国科学院声学研究所
Institute of Acoustics, CAS

无监督领域自适应

基于统计分布的领域自适应

度量两个领域的统计分布差异，通过训练使其最小化



跨域声纹识别： 无监督自适应

基于统计分布的领域自适应

相关对齐（Correlation Align，CORAL）

$$\min_{X_s,Y_s} L_C + \min_{F_s,F_t} \max_{D} \lambda L(D,F_s,F_t) + \sigma L_{DM}(F_s,F_t)$$

$$\min_{F_s,F_t} L_{DM}(F_s,F_t) = \frac{1}{4d^2} \| C_s - C_t \|_F^2$$

$$C_s = \frac{1}{N_s - 1}(F_s^T F_s - \frac{1}{N_s}(1^T F_s)^T (1^T F_s))$$

$$C_t = \frac{1}{N_t - 1}(F_t^T F_t - \frac{1}{N_t}(1^T F_t)^T (1^T F_t))$$

跨域声纹识别： 无监督自适应

中国科学院声学研究所
Institute of Acoustics, CAS

- 基于统计分布的领域自适应

最大均值差异（Maximum Mean Discrepancy，MMD）

$$\min_{X_s,Y_s} L_C + \min_{F_s,F_t} \max_{D} \lambda L(D,F_s,F_t) + \sigma \hat{d}_H(P_s,P_t)$$

实际应用中估算方式

$$\hat{d}_H(P_s,P_t) = \| \frac{1}{N_s}\sum \phi(x_i^s) - \frac{1}{N_t}\sum \phi(x_j^t) \|_H^2$$

$$= \frac{1}{N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s} k(x_i^s,x_j^s) + \frac{1}{N_t^2}\sum_{i=1}^{N_t}\sum_{j=1}^{N_t} k(x_i^t,x_j^t)$$

$$- \frac{2}{N_s N_t}\sum_{i=1}^{N_s}\sum_{j=1}^{N_t} k(x_i^s,x_j^t)$$

跨域声纹识别： 无监督自适应

中国科学院声学研究所
Institute of Acoustics, CAS

基于统计分布的领域自适应

实验1：跨设备

	数据集	领域	标注
训练	VoxCeleb 训练集	源	有
	SITW 开发集	目标	无
测试	SITW 测试集	目标	\

实验结果

测试场景	core-core	Core-multi	Assist-core	Assist-multi
基线	7.217	9.358	9.282	10.972
领域自适应	**6.670**	**8.950**	**8.783**	**10.369**

SITW有四种测试场景，core表示注册或测试语音中只有一个说话人，assist和multi分别表示注册和测试语音中有多个说话人

实验分析

基于统计分布的领域自适应

实验2：远近场

	数据集	领域	标注
训练	AISHELL-wakeup 训练集 近场	源	有
	AISHELL-wakeup 训练集 远场	目标	无
测试	AISHELL-wakeup 近场注册 远场测试	目标	\

实验结果

训练策略	EER（%）
CE	17.860
CE – Domain	16.164
CE + CORAL	16.322
CE + MMD	16.411
CE – domain + CORAL	**15.816**

实验分析

中国科学院声学研究所
Institute of Acoustics, CAS

- 基于统计分布的领域自适应
  - 实验3：中英文

	数据集	领域	标注
训练	VoxCeleb 训练集	源	有
	CN-Celeb 训练集	目标	无
测试	CN-Celeb 测试集	目标	\

  - 实验结果

训练策略	EER（%）
CE	17.188
CE – Domain	17.271
CE + CORAL	**15.816**
CE + MMD	15.832
CE – domain + CORAL	15.854

相关文章

- Wenjie Li, Pengyuan Zhang, Yonghong Yan, TEnet: target speaker extraction network with accumulated speaker embedding for automatic speech recognition, Electronics Letters, 2019，55(14)，816.
- Wenjie Li, Pengyuan Zhang, Yonghong Yan, Target speaker recovery and recognition network with average x-vector and global training, Interspeech 2019,3233.
- Hangting Chen, Pengyuan Zhang, Qian Shi and Zuozhen Liu. "Improved Guided Source Separation Integrated with a Strong Back-end for the CHiME-6 Dinner Party Scenario." Proc. Interspeech 2020 (2020): 334-338.
- Xueshuai Zhang, Wenchao Wang and Pengyuan Zhang. "Speaker Diarization System based on DPCA Algorithm For Fearless Steps Challenge Phase-2." Proc. Interspeech 2020 (2020): 2602-2606.
- Zhigao Chen, Xiaoxiao Miao, Runqiu Xiao and Wenchao Wang. "Cross-domain speaker recognition using domain adversarial siamese network with a domain discriminator." Electronics Letters (2020).

谢谢！