清华大学 Tsinghua University

新疆大学 Xinjiang University

声纹识别技术前沿与团队工作进展

清华大学 新疆大学

何 亮 副研究员

2020年11月21日

目录

contents

http://www.auroralab.top

清華大學 Tsinghua University

恶意攻击

白盒/黑盒

Antispoof

鉴别真假语音

录音回放

真实/欺骗

真/假

说话人识别/声纹识别

判断两段语音是否属于同一个说话人

段/流

说话人标记

语音段到语音流

多维属性

段

流

语种/性别/情绪

语种、性别与情绪等辅助或联合识别

语音内容

活动语音/连续语音识别等辅助或联合识别

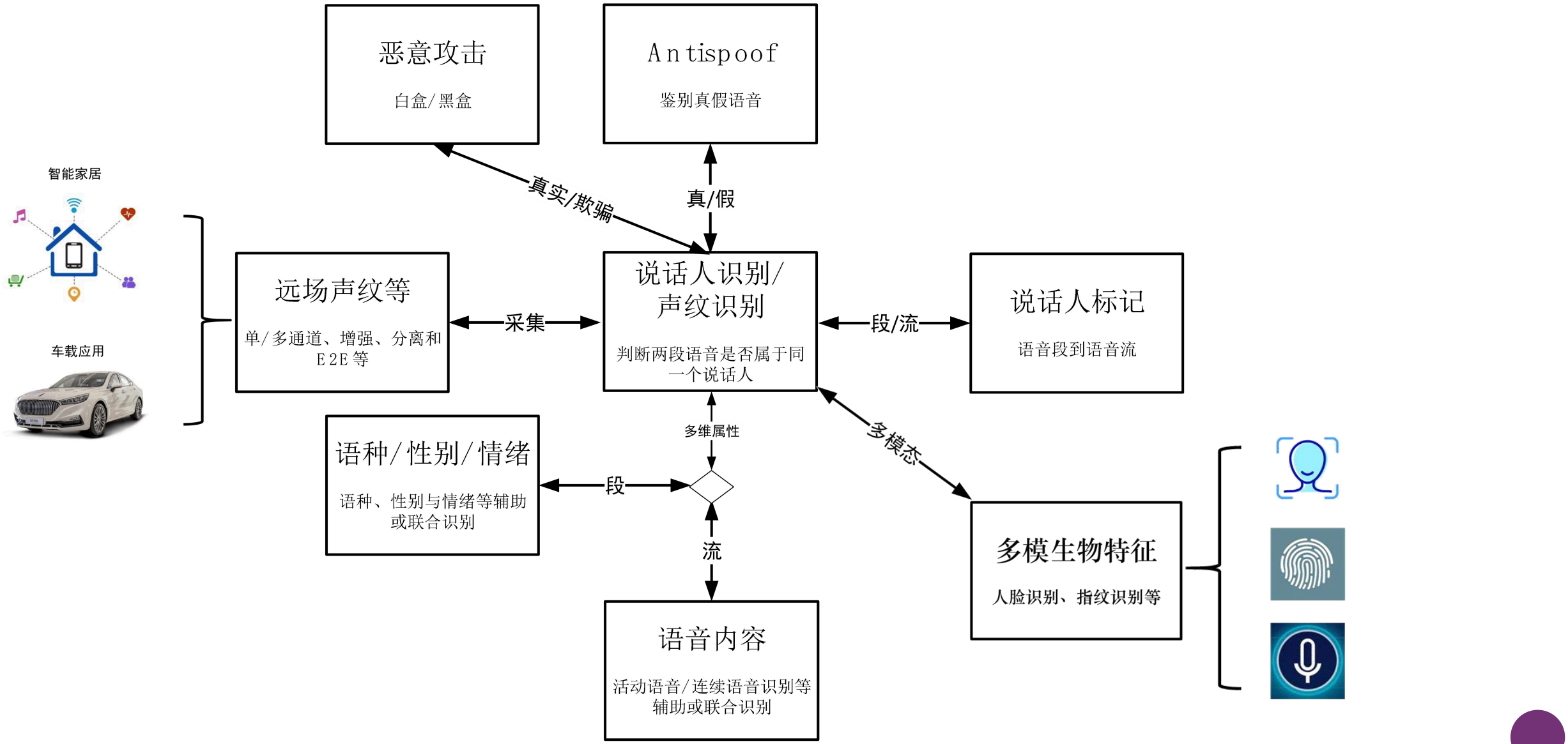
语音合成或转换

声纹与内容联合识别

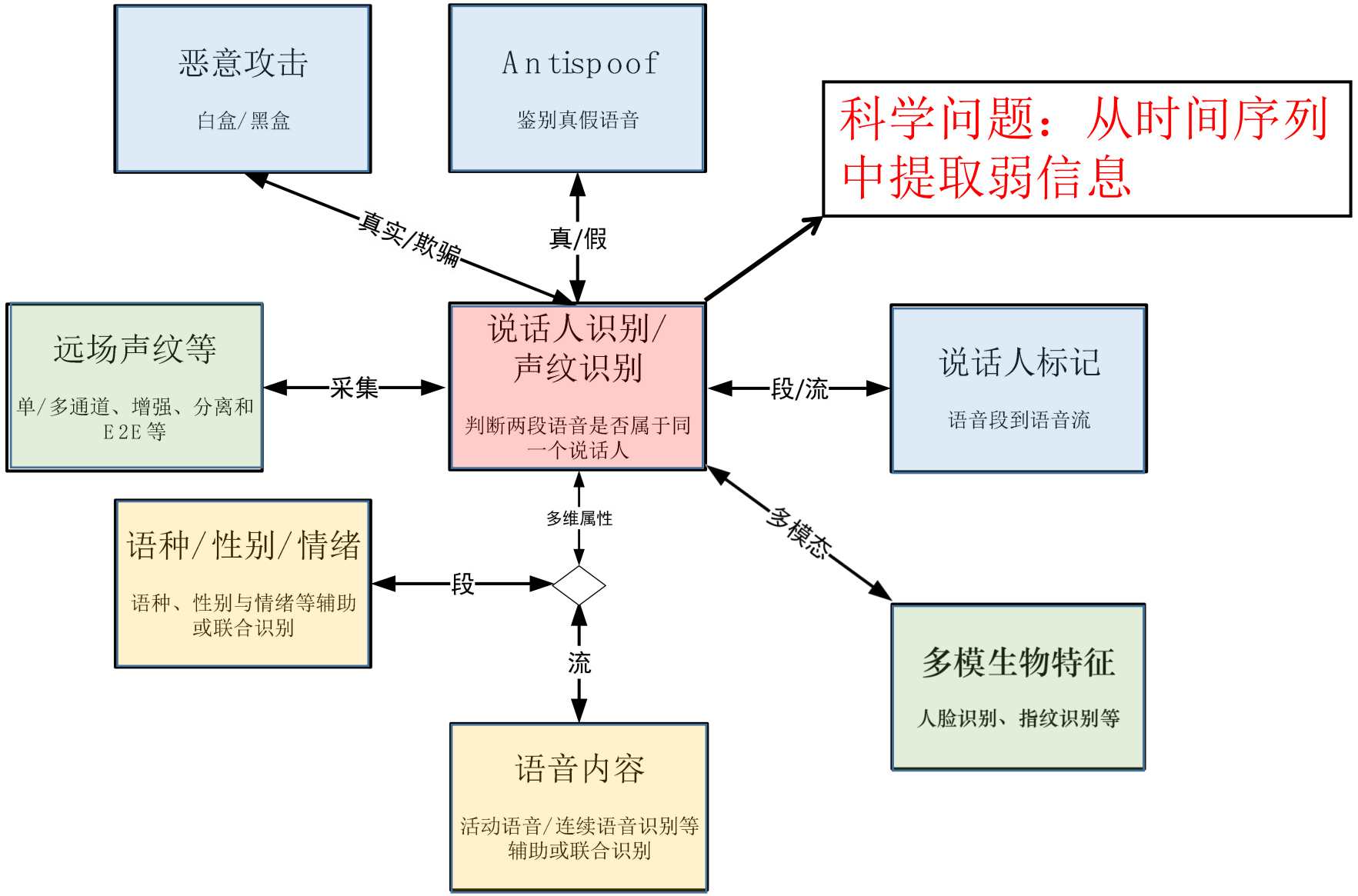
今天有声纹研讨会

何亮

http://www.auroralab.top



http://www.auroralab.top



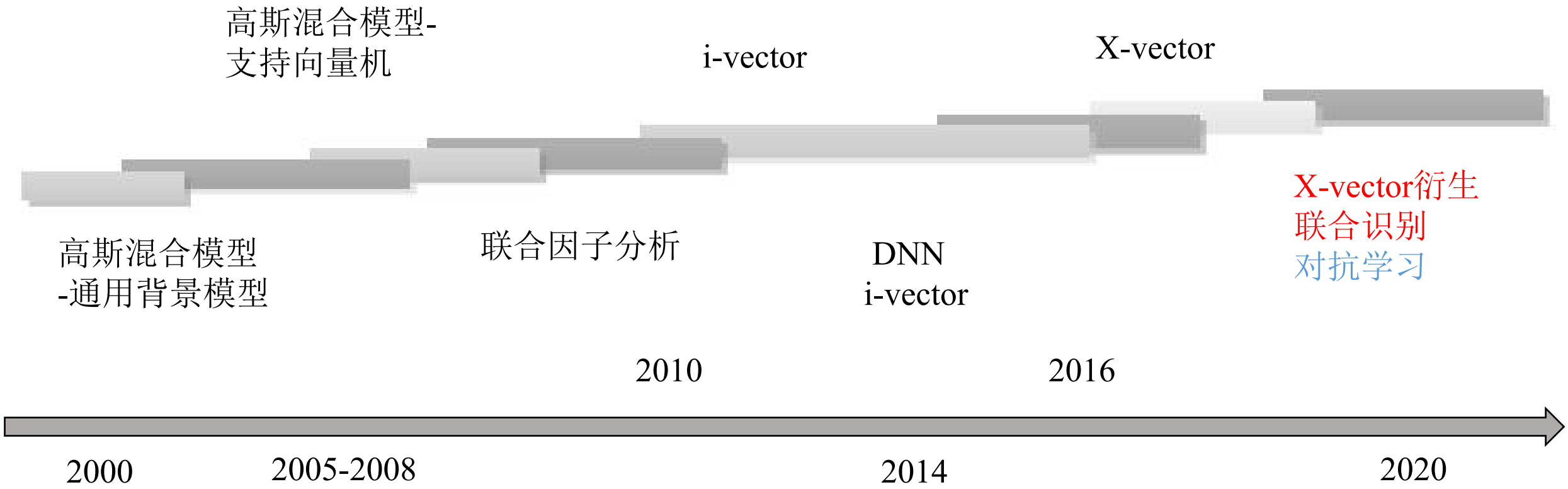
清華大學 Tsinghua University

新疆大學 Xinjiang University

一、技术前沿

深度神经网络的改进、联合识别及对抗学习

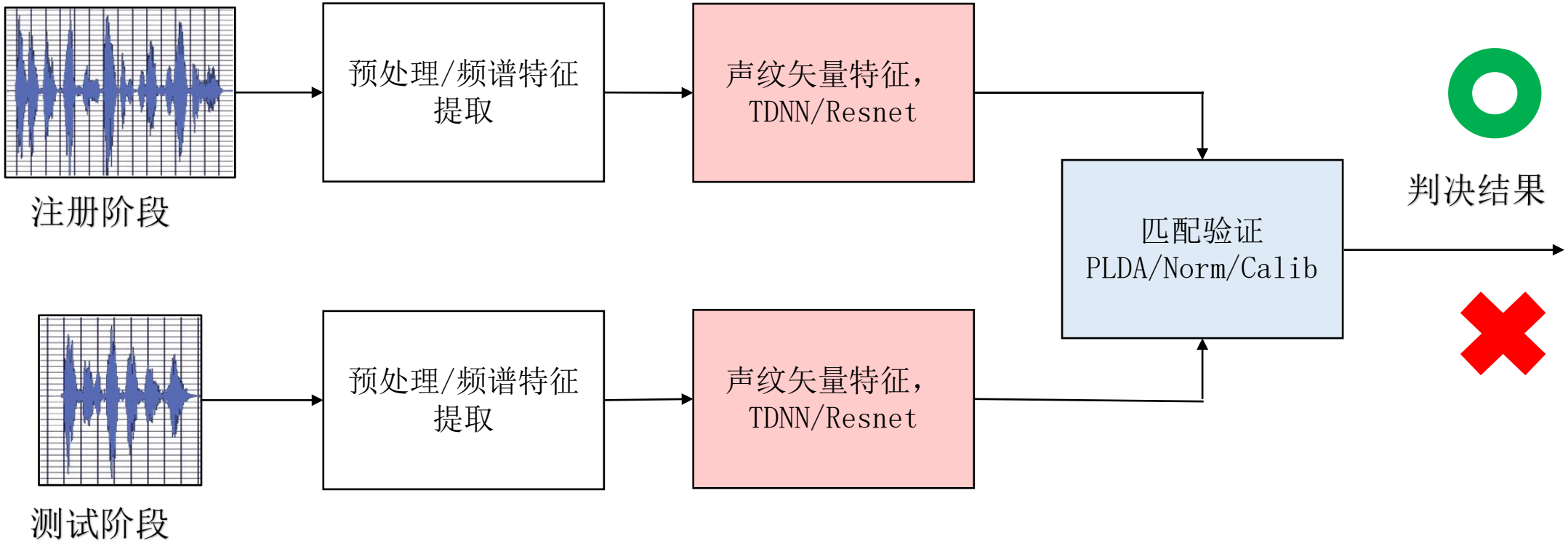
历史回顾



主要参考：NIST SRE19/20、VoxCeleb
IEEE ASLP、IEEE SPL、SC、CSL、EURASIP ASMP、IS20、ICASSP20、Odyssey20和arXiv

技术前沿

- 声纹识别基本框架



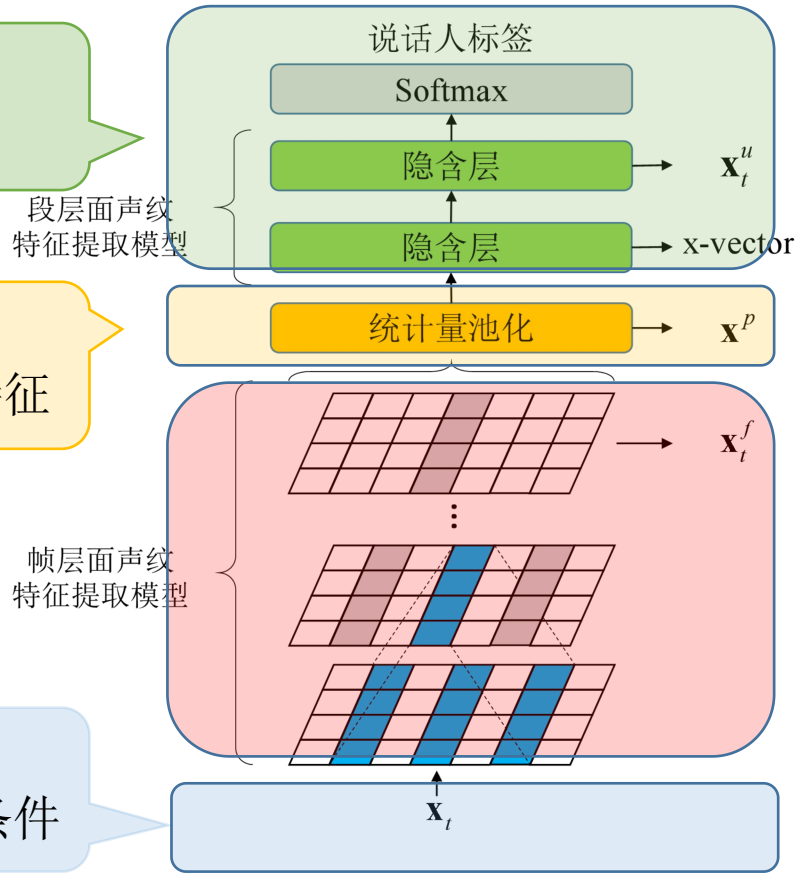
技术前沿

- X-vector分析

损失函数
目标：最小类内最大类间

池化方法
目标：凸显段层面声纹特征

数据扩充
目标：尽可能覆盖训练测试条件



Xvector

段层面声纹特征

声纹特征矢量

池化结果

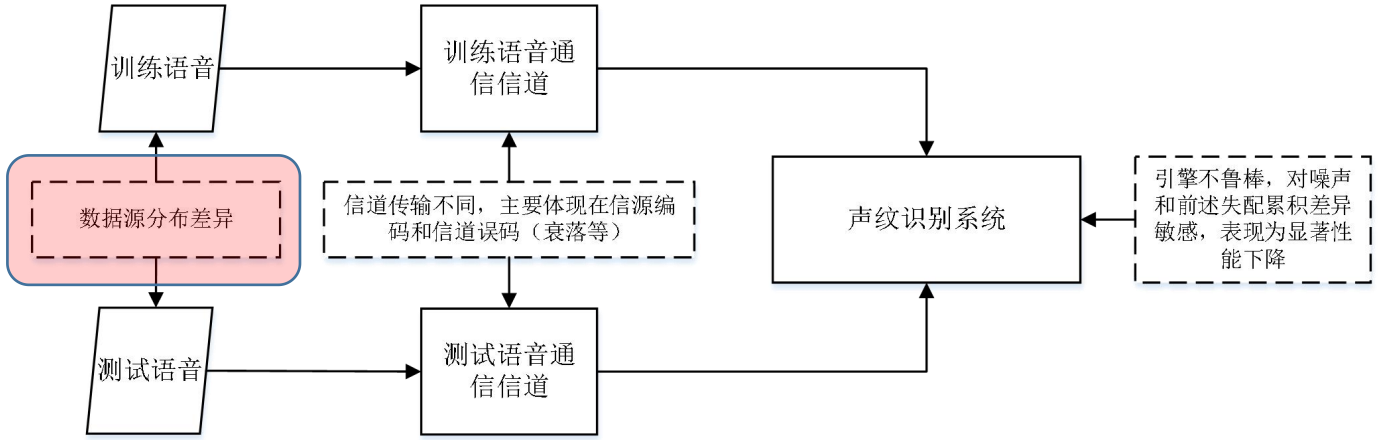
帧层面声纹特征

深度神经网络结构
目标：凸显帧层面声纹特征

Cosine (E2E) vs PLDA (Domain)
抗领域失配、分数校准

技术前沿

- 数据扩充：从抗信道失配到抗说话方式失配
  - 1. 根据Xvector论文/Kaldi脚本扩充
  - 2. 随机删除
  - 3. Speaking style：朗读、对话和陈述等



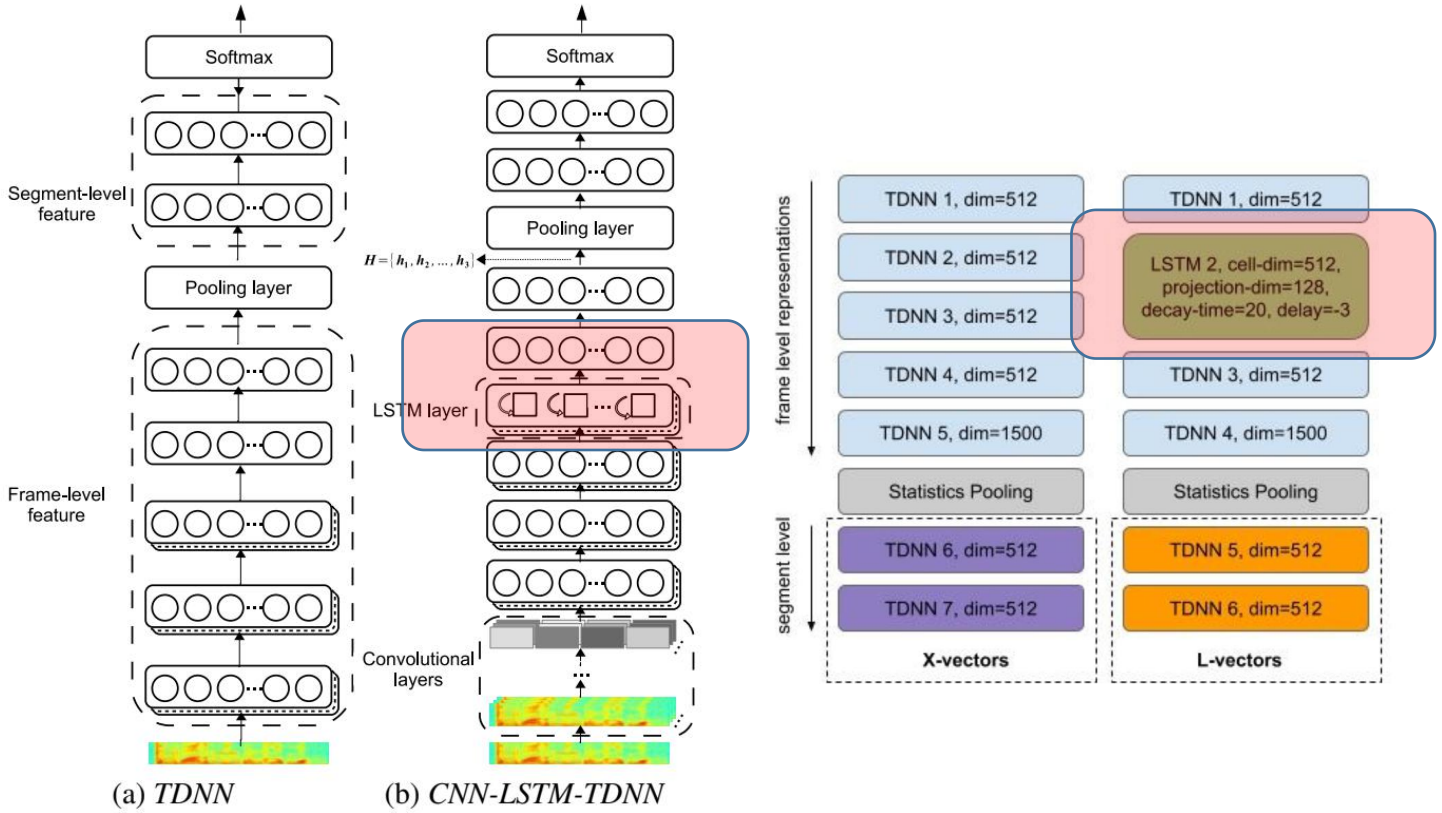
UCLA和JHU，针对speaker-style variability，基于交叉熵，提出基于变帧速率的数据增强方法

NIT，通过VC方法生成语音，解决失配问题

1. A. Afshan, J. Guo, S. J. Park, V. Ravi, A. McCree, and A. Alwan, “Variable Frame Rate-Based Data Augmentation to Handle Speaking-Style Variability for Automatic Speaker Verification,” in *Interspeech 2020*, Oct. 2020, pp. 4318–4322.
2. S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, “In-Domain and Out-of-Domain Data Augmentation to Improve Children’s Speaker Verification System in Limited Data Scenario,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7554–7558.

技术前沿

- 深度神经网络结构
  - 评测验证的技术：
    - SRE19：(F-E) TDNN-SE，LSTMP/OPGRU
    - VoxCeleb20：ECAPA-TDNN
  - 设计出发点：时序、多尺度和注意力机制



1. R. Li, D. Chen, and W. Zhang, "Voiceai Systems to NIST Sre19 Evaluation: Robust Speaker Recognition on Conversational Telephone Speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6459–6463.
2. J. Thienpondt, B. Desplanques, and K. Demuynck, "The IDLAB VoxCeleb Speaker Recognition Challenge 2020 System Description," *arXiv:2010.12468 [cs, eess]*, Oct. 2020, Accessed: Nov. 12, 2020. [Online]. Available: http://arxiv.org/abs/2010.12468.

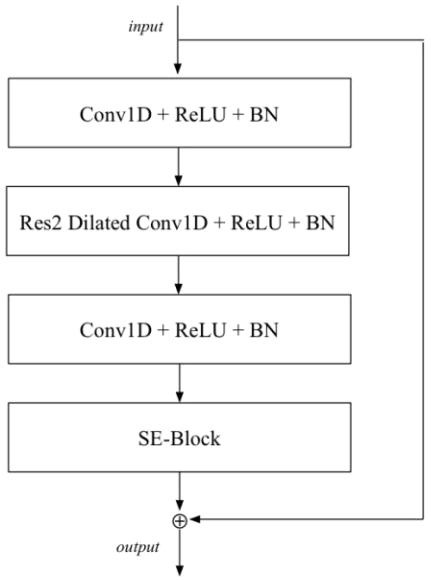
技术前沿

- 深度神经网络结构
  - ECAPA-TDNN：19% Vox

Res2Net：层内多尺度

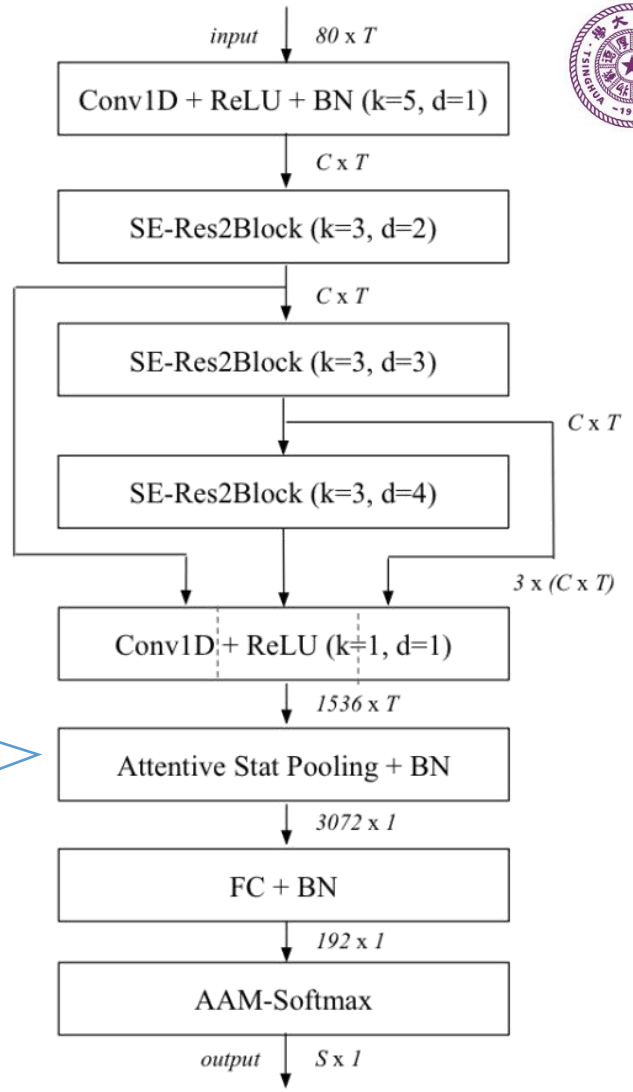
SE Block：信道注意力

残差设计



跨层聚合，层面的多尺度

注意力统计池化：时间的选择性

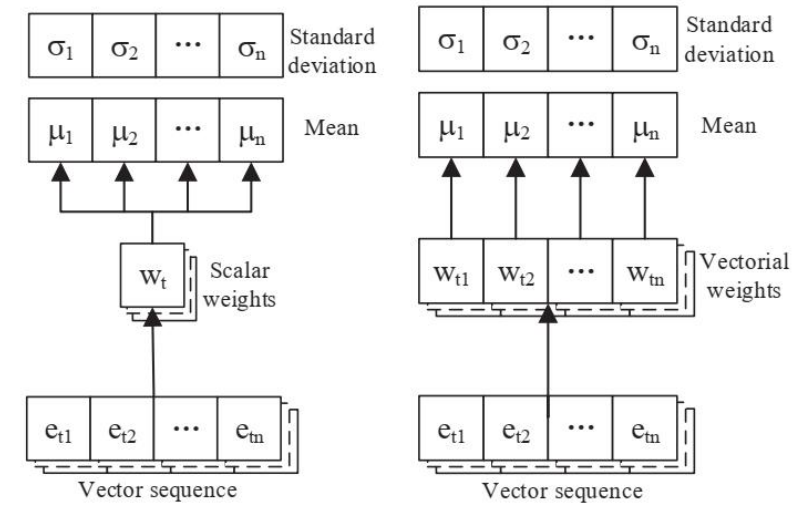


时序信息？

1. B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” *Interspeech 2020*, pp. 3830–3834, Oct. 2020, doi: 10.21437/Interspeech.2020-2650.

技术前沿

- 池化方法：从点到向量
  - Attentive, Multi-head -> Vector-based



(a) Attentive statistics pooling. (b) Vector-based attentive pooling.

Table 3: *Experimental results on VoxCeleb.* ***Boldface*** *values are the best results.*

Embedding	EER(%)	DCF10^{-2}	DCF10^{-3}
i-vector [1]	5.657	0.5016	0.6593
statistics [3]	2.556	0.3079	0.5582
self multi-head [20]	2.709	0.2804	**0.4032**
attentive statistics [18]	2.593	0.2947	0.4322
self-attentive(1) [19]	2.667	0.3002	0.4307
self-attentive(2) [19]	2.773	0.2940	0.4877
self-attentive(5) [19]	2.635	0.2887	0.4041
vector-based attentive(1)	2.582	0.2894	0.5126
vector-based attentive(2)	**2.466**	**0.2726**	0.4286
vector-based attentive(3)	2.641	0.3070	0.4107

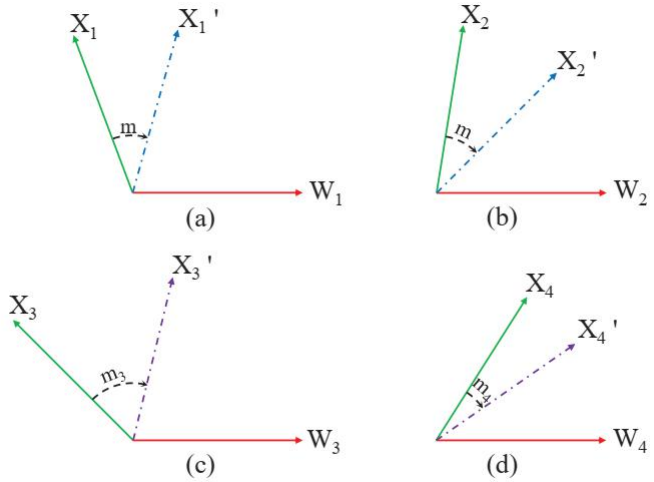
1. Y. Wu, C. Guo, H. Gao, X. Hou, and J. Xu, "Vector-Based Attentive Pooling for Text-Independent Speaker Verification," in *Interspeech 2020*, Oct. 2020, pp. 936–940, doi: 10.21437/Interspeech.2020-1422.

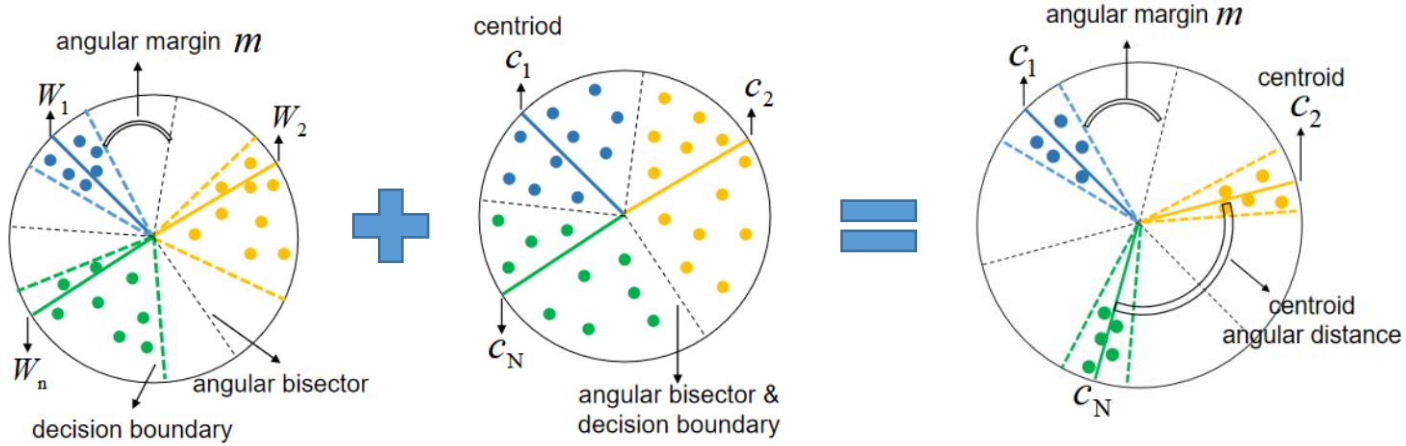
技术前沿

Table 1: *Speaker verification EER (%) and speaker identification accuracy (%).*

Loss	EER	Accuracy
Softmax	10.43	80.51
Triplet Loss (*cosine*, $\alpha = 0.1$)	8.41	82.62
GE2E Loss	8.30	83.74
AM-Softmax ($m = 0.0$)	11.36	79.12
AM-Softmax ($m = 0.3$)	9.85	81.53
AM-Softmax ($m = 0.4$)	8.01	84.61
AM-Softmax ($m = 0.5$)	7.38	85.28
AM-Centroid Loss ($m = 0.3$)	7.72	85.14
AM-Centroid Loss ($m = 0.4$)	6.59	86.37
AM-Centroid Loss ($m = 0.5$)	6.14	86.51

- 损失函数：静态间距到动态间距，点到中心
  - DAM、GE2E和AMCL

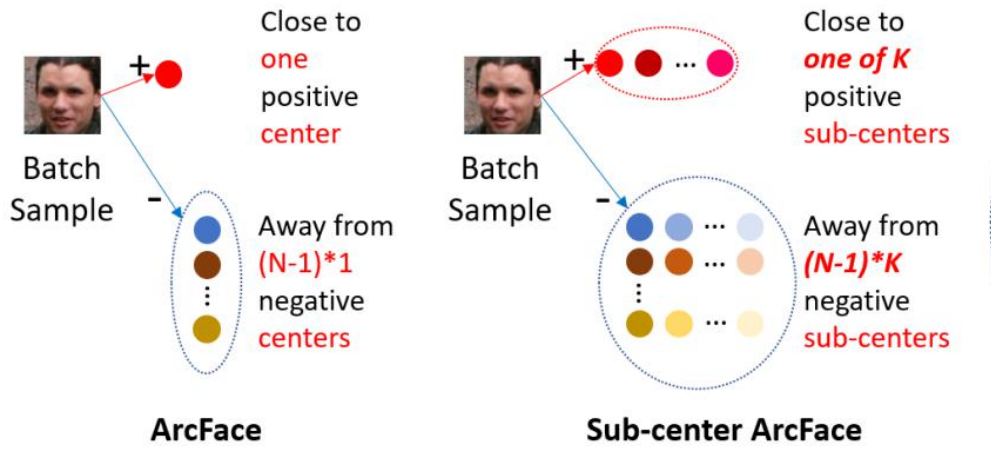




1. D. Zhou *et al.*, "Dynamic Margin Softmax Loss for Speaker Verification," in *Interspeech 2020*, Oct. 2020, pp. 3800–3804.
2. L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Apr. 2018, pp. 4879–4883, doi: 10.1109/ICASSP.2018.8462665.
3. Y. Wei, J. Du, and H. Liu, "Angular Margin Centroid Loss for Text-Independent Speaker Recognition," in *Interspeech 2020*, Oct. 2020, pp. 3820–3824, doi: 10.21437/Interspeech.2020-2538.

技术前沿

- 损失函数：单中心到多中心



(a) ArcFace vs. Sub-center ArcFace

Table 2: *Evaluation of all fine-tuned systems in the final fusion submission of the closed track on VoxSRC-20 validation set.*

Architecture	Variant	EER(%)	$\text{MinDCF}_{0.01}$
ECAPA-TDNN	Baseline	2.89	0.2274
ECAPA-TDNN	Tanh in CAS	2.86	0.2274
ECAPA-TDNN	BLSTM	2.88	0.2360
ECAPA-TDNN	256-dim emb.	3.15	0.2578
ECAPA-TDNN	FBANK60	2.92	0.2389
ECAPA-TDNN	SC-AAM & DD	2.83	0.2298
ResNet34	SE-blocks	3.03	0.2605
SE-ResNet34	CAS	2.89	0.2306
SE-ResNet34	SC-AAM	2.98	0.2437
SE-ResNet34	SC-AAM & CAS	**2.70**	**0.2215**
Fusion		2.41	0.1901
Fusion + QMFs		**2.16**	**0.1795**

1 J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020

2. J. Thienpondt, B. Desplanques, and K. Demuynck, "The IDLAB VoxCeleb Speaker Recognition Challenge 2020 System Description," *arXiv:2010.12468 [cs, eess]*, Oct. 2020, Accessed: Nov. 12, 2020. [Online]. Available: http://arxiv.org/abs/2010.12468.

技术前沿

- 后端处理：E2E的效果并不鲁棒，倾向于将PLDA融入E2E
  - 领域失配：COREL/PLDA的插值
  - Neural PLDA, Noisy Labels

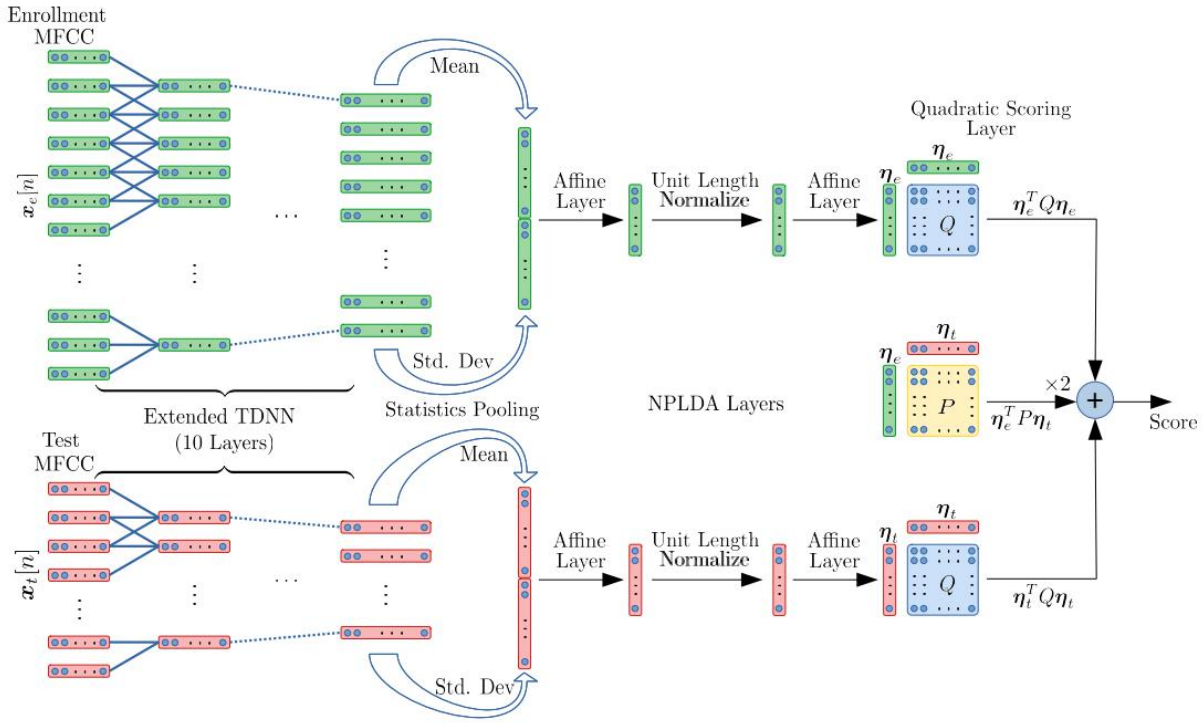


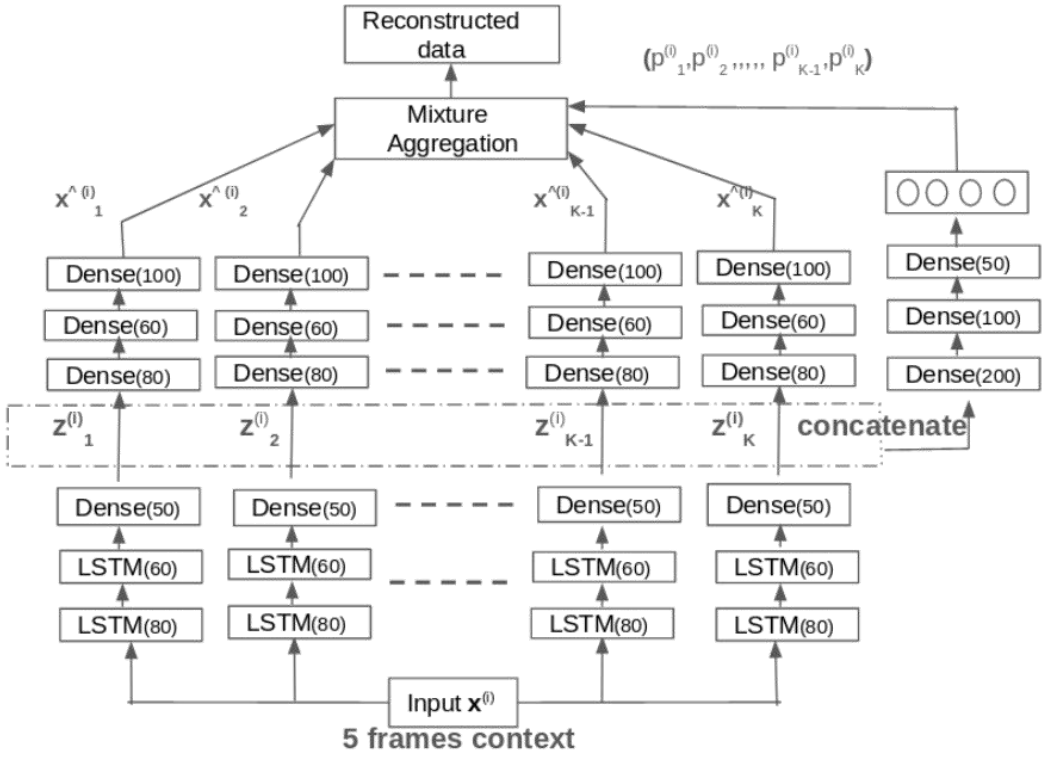
Figure 1: *End-to-End x-vector NPLDA architecture for Speaker Verification.*

[1] S. Ramoji, P. Krishnan, and S. Ganapathy, "Neural PLDA Modeling for End-to-End Speaker Verification," in *Interspeech 2020*, Oct. 2020, pp. 4333–4337, doi: 10.21437/Interspeech.2020-2699.

[2] B. J. Borgstrom and P. Torres-Carrasquillo, "Bayesian Estimation of Plda with Noisy Training Labels, with Applications to Speaker Verification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7594–7598,

技术前沿

- 语音内容联合识别
  - 真实标签->伪标签



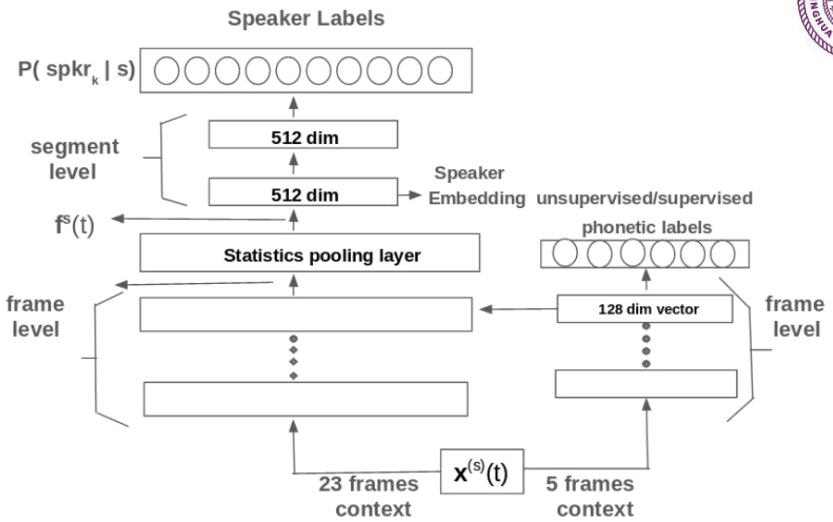


Table 1. Performance evaluation on NIST SRE-10 condition-5

System	EER(%) (minDCF08)		
	core-core	core-10s	10s-10s
i-vector	2.31 (0.0129)	6.376 (0.0353)	12.09 (0.0570)
x-vector	1.82 (0.0104)	4.866 (0.0256)	9.89 (0.0513)
c-vector	1.33 (0.0073)	4.19 (0.0235)	8.24 (0.0439)
uc-vector	1.60 (0.0084)	4.69 (0.0248)	8.24 (0.0449)
Score fusion			
i+c	1.395 (0.0076)	3.691 (0.0212)	7.509 (0.0398)
i+uc	1.395 (0.0077)	4.027 (0.0202)	6.777 (0.0387)
x+c	1.367 (0.0073)	3.523 (0.0215)	7.326 (0.0391)
x+uc	1.451 (0.0076)	3.859 (0.0208)	6.96 (0.0376)
c+uc	1.283 (0.0069)	3.523 (0.0194)	6.41 (0.0348)
i+x+c	1.353 (0.0074)	3.356 (0.0197)	6.777 (0.0359)
i+x+uc	1.367 (0.0075)	3.188 (0.0189)	6.777 (0.0354)
i+x+c+uc	**1.241 (0.0068)**	**2.852 (0.0182)**	**5.495 (0.0317)**

1. S. Sreekanth, S. M. Rafi B, K. Sri Rama Murty, and S. Bhati, "Speaker Embedding Extraction with Virtual Phonetic Information," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Ottawa, ON, Canada, Nov. 2019, pp. 1–5, doi: 10.1109/GlobalSIP45357.2019.8969551.


清華大學
Tsinghua University

新疆大學
Xinjiang University

二、团队工作

联合学习、对抗学习和潜在类别分析

多任务学习

- 动机：语音内容和说话人信息被听者共同感知，知悉一个维度的信息对另一个维度信息的识别与理解有显著提升
- 历史工作：Multi-task learning + Phonetic adaptation

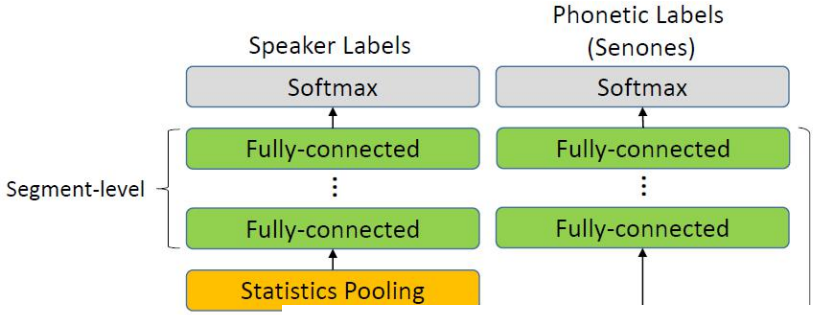


Table 6 Results of systems using hybrid multi-task learning with different configurations on VoxCeleb

	EER(%)	minDCF08	minDCF10
x-vector	2.68	0.0144	0.4645
x-vector-mt (1-layer sharing)	2.58	0.0132	0.4027
x-vector-mt (2-layer sharing)	2.73	0.0145	0.3977
x-vector-mt (3-layer sharing)	2.83	0.0151	0.4700
x-vector-mt (4-layer sharing)	2.92	0.0151	0.5001

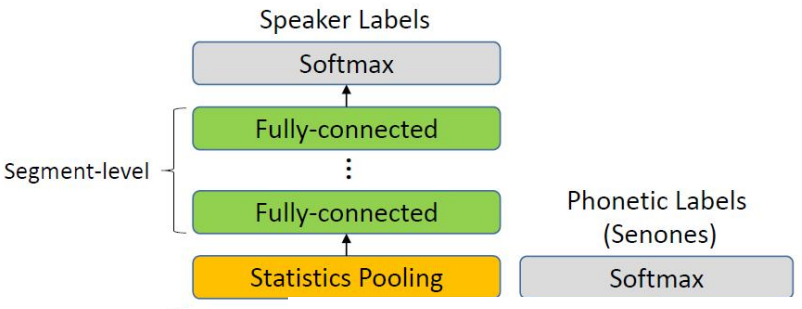
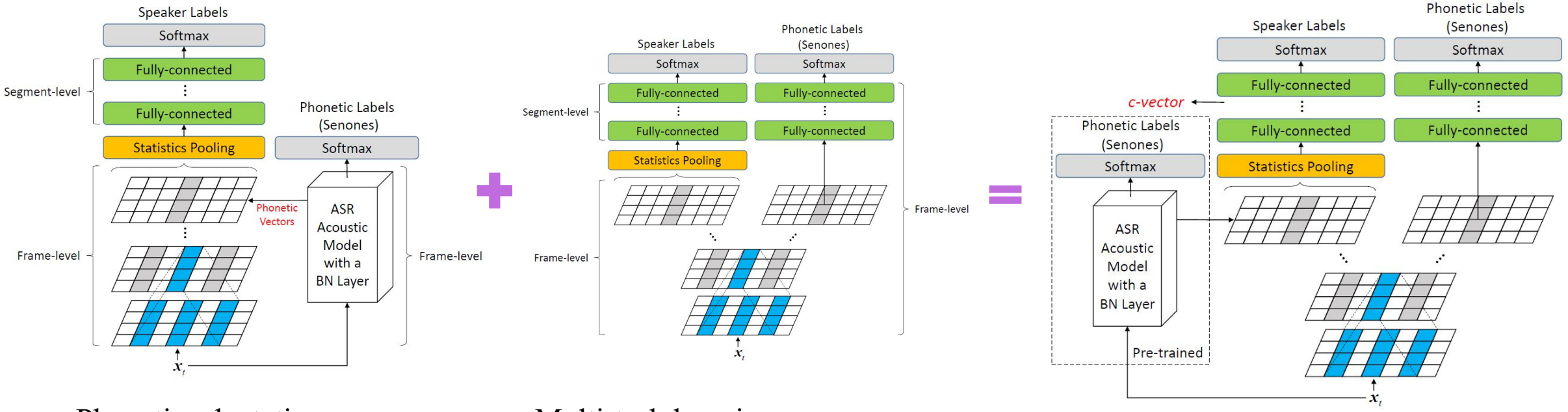


Table 4 Phonetic adaptation results on VoxCeleb

	EER(%)	minDCF08	minDCF10
x-vector	2.68	0.0144	0.4645
x-vector-pa ($c = 0$)	2.52	0.0137	0.4111
x-vector-pa ($c = 0.1$)	2.26	0.0126	0.3651
x-vector-pa ($c = 0.2$)	2.24	0.0132	0.3289
x-vector-pa ($c = 0.3$)	2.32	0.0124	0.3533
x-vector-pa ($c = 0.4$)	2.32	0.0126	0.3531

多任务学习

- 改进
  - MTL和PA对音素的利用层面不同
  - Factorized-TDNN



Phonetic adaptation　　Multi-task learning　　FC-vector

多任务学习

- 实验结果
  - VoxCeleb和SRE18

Table 13 Summary of results obtained with different systems on VoxCeleb

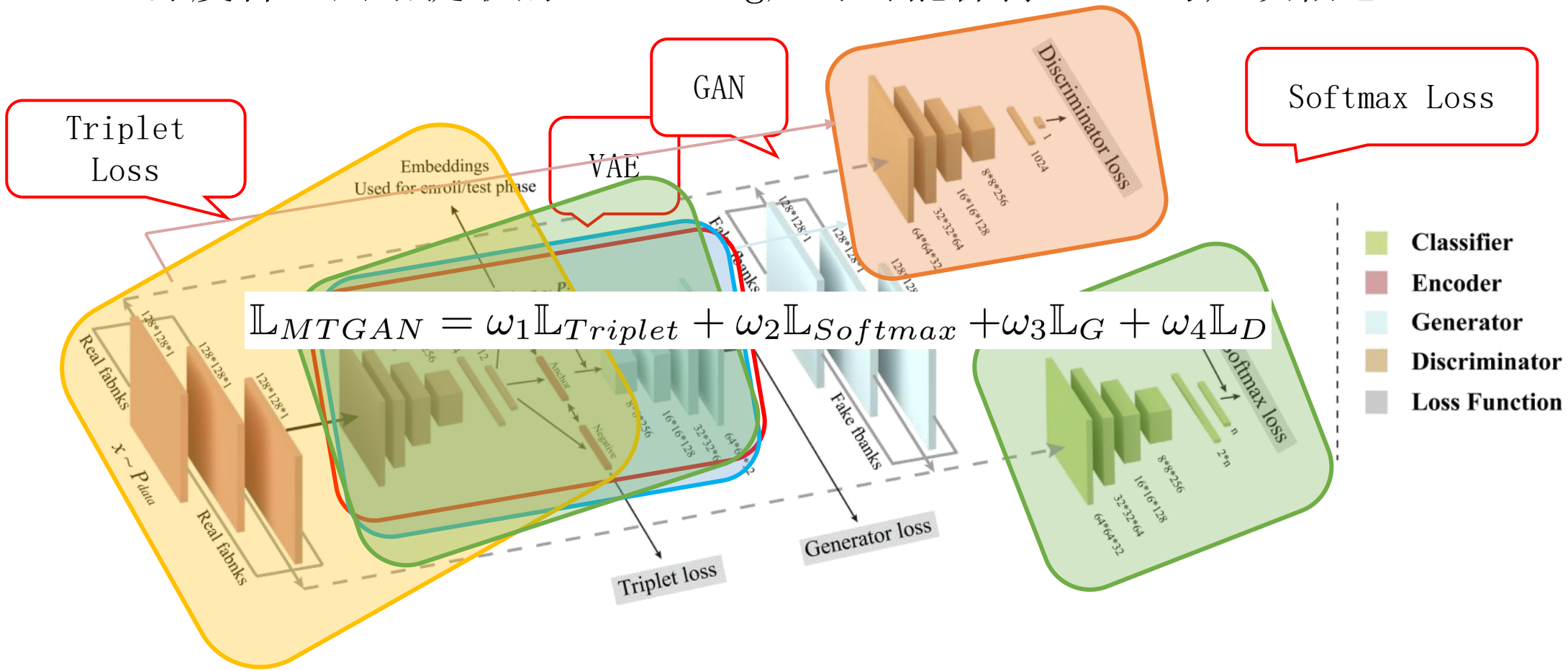
	EER(%)	minDCF08	minDCF10
x-vector	2.68	0.0144	0.4645
x-vector-pa ($c = 0$)	2.52	0.0137	0.4111
x-vector-pa ($c = 0.2$)	2.24	0.0132	0.3289
x-vector-mt (1-layer sharing)	2.58	0.0132	0.4027
sc-vector (1-layer sharing)	2.52	0.0142	0.4333
c-vector ($c = 0.2$ + 1-layer sharing)	2.18	0.0129	0.2994

System	SRE18 Dev		SRE18 Eval	
	EER(%)	minDCF18	EER(%)	minDCF18
x-vector	7.43	0.484	7.80	0.550
EF-TDNN	6.35	0.452	7.09	0.500
c-vector (1-layer sharing)	6.86	0.489	7.09	0.512
fc-vector (1-layer sharing)	6.34	**0.418**	**6.85**	**0.492**
fc-vector (3-layer sharing)	6.59	0.433	7.07	0.509
fc-vector (5-layer sharing)	**6.26**	0.467	7.10	0.513

1. Liu Yi, Liang He, Jia Liu and Michael T. Johnson, "Introducing phonetic information to speaker embedding for speaker verification," Eurasip Journal on Audio, Speech, and Music Processing, vol. 2019, no. 1, 2019.
2. Tianyu Liang, Yi Liu, Can Xu, Xianwei Zhang and Liang He, "Combined Vector Based on Factorized Time-delay Neural Network for Text-Independent Speaker Recognition," Odyssey 2020 The Speaker and Language Recognition Workshop, 01-05, Nov 2020, Tokyo, Japan, 428-432.

对抗学习

- 动机：
  - 深度神经网络提取的embedding应尽可能保持Fbank与声纹信息



$$\mathbb{L}_{MTGAN} = \omega_1 \mathbb{L}_{Triplet} + \omega_2 \mathbb{L}_{Softmax} + \omega_3 \mathbb{L}_G + \omega_4 \mathbb{L}_D$$

对抗学习

清華大學 Tsinghua University

- 实验：

训练集：Librispeech（1252人，每人20段）
测试集：TIMIT（630人，3段注册，7段测试）

Table 1: *EER (%) and accuracy (%) of different systems*

Methods	Equal Error Rate	Accuracy
i-vector/Cosine	8.48%	81.92%
i-vector/PLDA	5.61%	85.78%
Softmax loss [3]	3.61%	88.23%
Triplet loss [4]	2.68%	90.45%
MTGAN	**1.81%**	**92.65%**

Table 2: *Ablation experiments with different conditions*

Conditions	EER	ACC	Convergence	
w/o GAN	2.04%	90.17%	60 epoch	1
w/o softmax loss	3.34%	88.63%	80 epoch	
w/o triplet loss	2.71%	89.51%	60 epoch	
MTGAN	**1.81%**	**92.65%**	100 epoch	
Random (#60)	3.13%	85.26%	550 epoch	2
Semi-hard (#60)	2.90%	88.73%	500 epoch	
Random (#600)	2.75%	90.03%	250 epoch	
Semi-hard (#600)	**2.68%**	**90.45%**	200 epoch	
1252 people	1.81%	92.65%	70 epoch	3
2484 people	**1.33%**	**94.27%**	100 epoch	

1 测试各个模块贡献　　2 比较采样方法的区别　　3 比较训练人数的区别

对抗学习

- 分析：模块复用，多策略强化学习内容；帧层面对抗学习有效

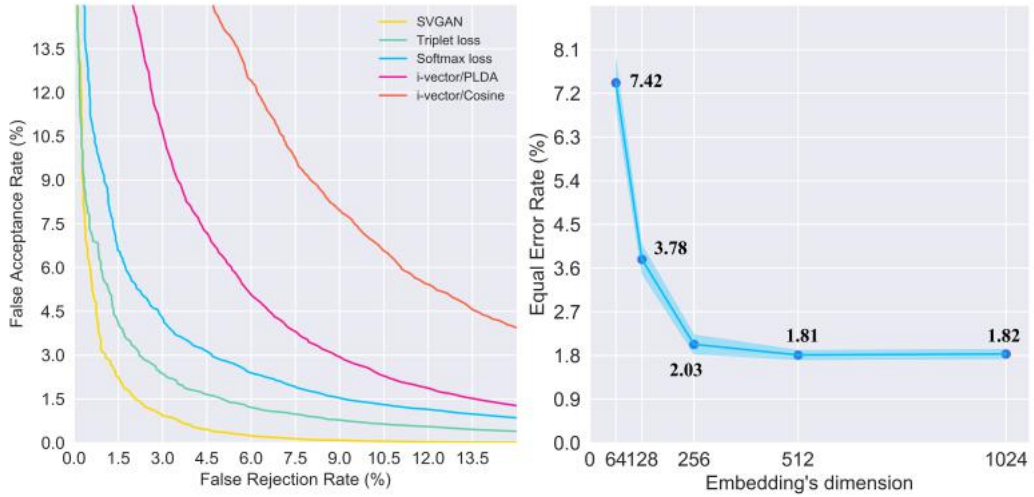


Figure 3: ***Left:*** *DET curves. Results of five methods are displayed. Two back-end methods are used for i-vector system.* ***Right:*** *EER (%) with different embedding dimensions. Five kinds of dimensions of the embeddings are displayed.*

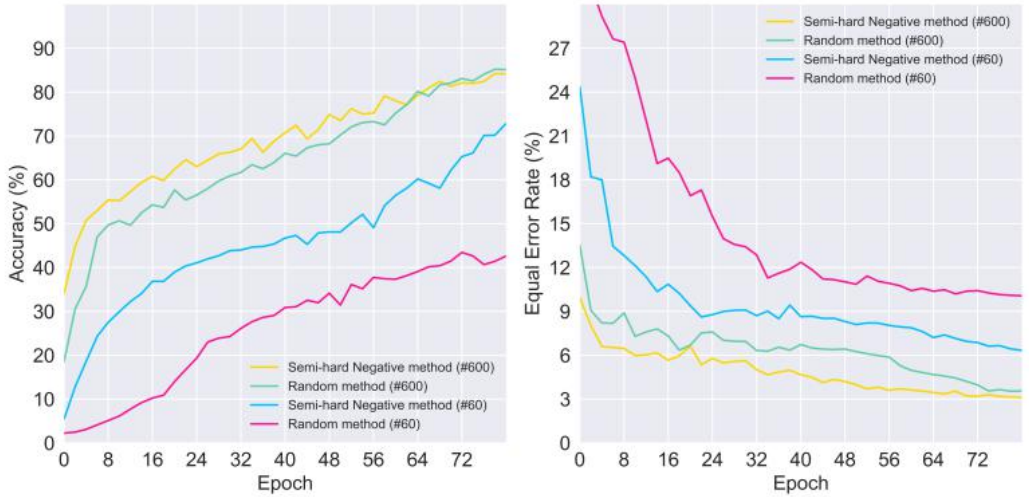


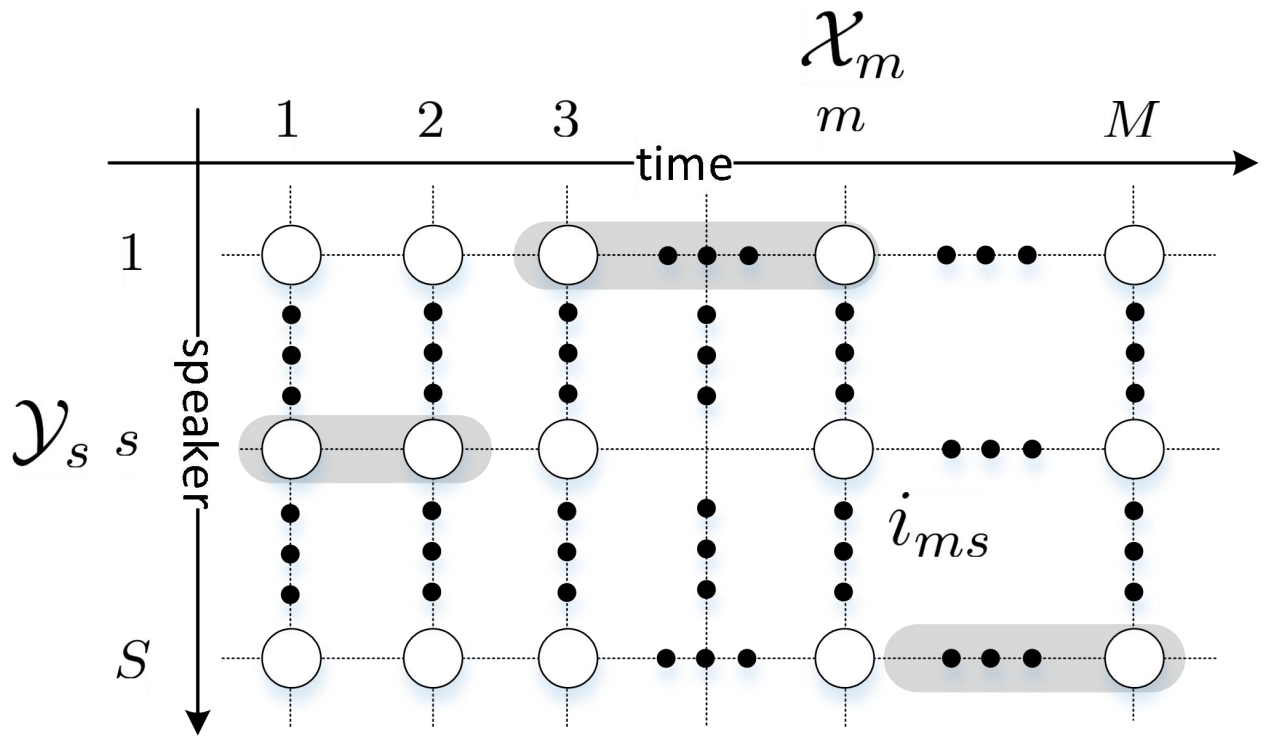
Figure 4: ***Changes of ACC (%) and EER (%) with the time of training.*** *We choose the first 80 epochs to show the trend.*

1 五种算法的DET曲线　　2 embedding维度的影响　　3 4 随训练轮数的ACC & EER 变化

1. Wenhao Ding and Liang He, "MTGAN: Speaker Verification through Multitasking Triplet Generative Adversarial Networks," INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, 2-6 September 2018, Hyderabad, 3633-3637.

说话人标记

- 如何看待说话人标记问题
  - 分割-聚类
  - 概率估计
- 问题的描述
  - 假设：
    - X是语音，Y是说话人模型
  - 已知：
    - 段长M与说话个数S
  - 估计：
    - 概率矩阵I，$i_{ms} = \begin{cases} 1, & \text{if segment } m \text{ belongs to the latent class } s \\ 0, & \text{if segment } m \text{ does not belong to the latent class } s \end{cases}$



说话人标记

- 目标函数：$$\max \log p(X, \mathcal{Y}, I) = \sum_{m=1}^{M} \log \sum_{s=1}^{S} p(\mathcal{X}_m, \mathcal{Y}_s, i_{ms})$$

  $$\text{s.t. S speakers}$$

- 困难：Y与I是隐含变量

- 解决方案：$$\max \sum_{m=1}^{M} \log \sum_{s=1}^{S} p(\mathcal{X}_m, \mathcal{Y}_s, i_{ms})$$
  $$= \sum_{m=1}^{M} \log \sum_{s=1}^{S} p(\mathcal{X}_m, \mathcal{Y}_s) p(i_{ms} | \mathcal{X}_m, \mathcal{Y}_s)$$
  $$= \sum_{m=1}^{M} \log \sum_{s=1}^{S} p(\mathcal{X}_m, \mathcal{Y}_s) q_{ms}$$
  $$\text{s.t.} \sum_{s=1}^{S} q_{ms} = 1, 0 \le q_{ms} \le 1$$

- 分析：
  - 我们关心的是：$p(i_{ms} | X_m, Y_s)$
  - X_m是以m时刻为中心的语音片段，是已知确定的
  - Y_s取决于说话人统计建模方法，是未知变化的
  - $p(i_{ms} | X_m, Y_s)$正比于$p(i_{ms}, Y_s | X_m)$

说话人标记

- 步骤1：通过引入辅助分布A和Jensen不等式，有

$$\sum_{m=1}^{M}\log\sum_{s=1}^{S}a_{ms}\frac{q_{ms}p(\mathcal{X}_m,\mathcal{Y}_s)}{a_{ms}}$$
$$\geq\sum_{m=1}^{M}\sum_{s=1}^{S}a_{ms}\log\frac{q_{ms}p(\mathcal{X}_m,\mathcal{Y}_s)}{a_{ms}}$$

其中

$$a_{ms}=\frac{q_{ms}p(\mathcal{X}_m,\mathcal{Y}_s)}{\sum_{s'=1}^{S}q_{ms'}p(\mathcal{X}_m,\mathcal{Y}_{s'})}$$

故而，使用a更新q

$$q_{ms}=a_{ms}$$

- 步骤2：

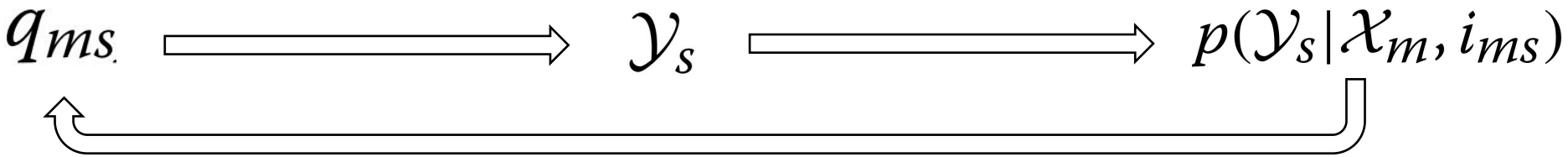
$$\max\sum_{m=1}^{M}\log\sum_{s=1}^{S}p(\mathcal{X}_m,\mathcal{Y}_s,i_{ms})$$
$$=\sum_{m=1}^{M}\log\sum_{s=1}^{S}p(i_{ms})p(\mathcal{X}_m,\mathcal{Y}_s|i_{ms})$$
$$=\sum_{m=1}^{M}\log\sum_{s=1}^{S}q_{ms}p(\mathcal{Y}_s)p(\mathcal{X}_m|\mathcal{Y}_s,i_{ms})$$

给定语音X和概率矩阵Q，估计说话人模型Y

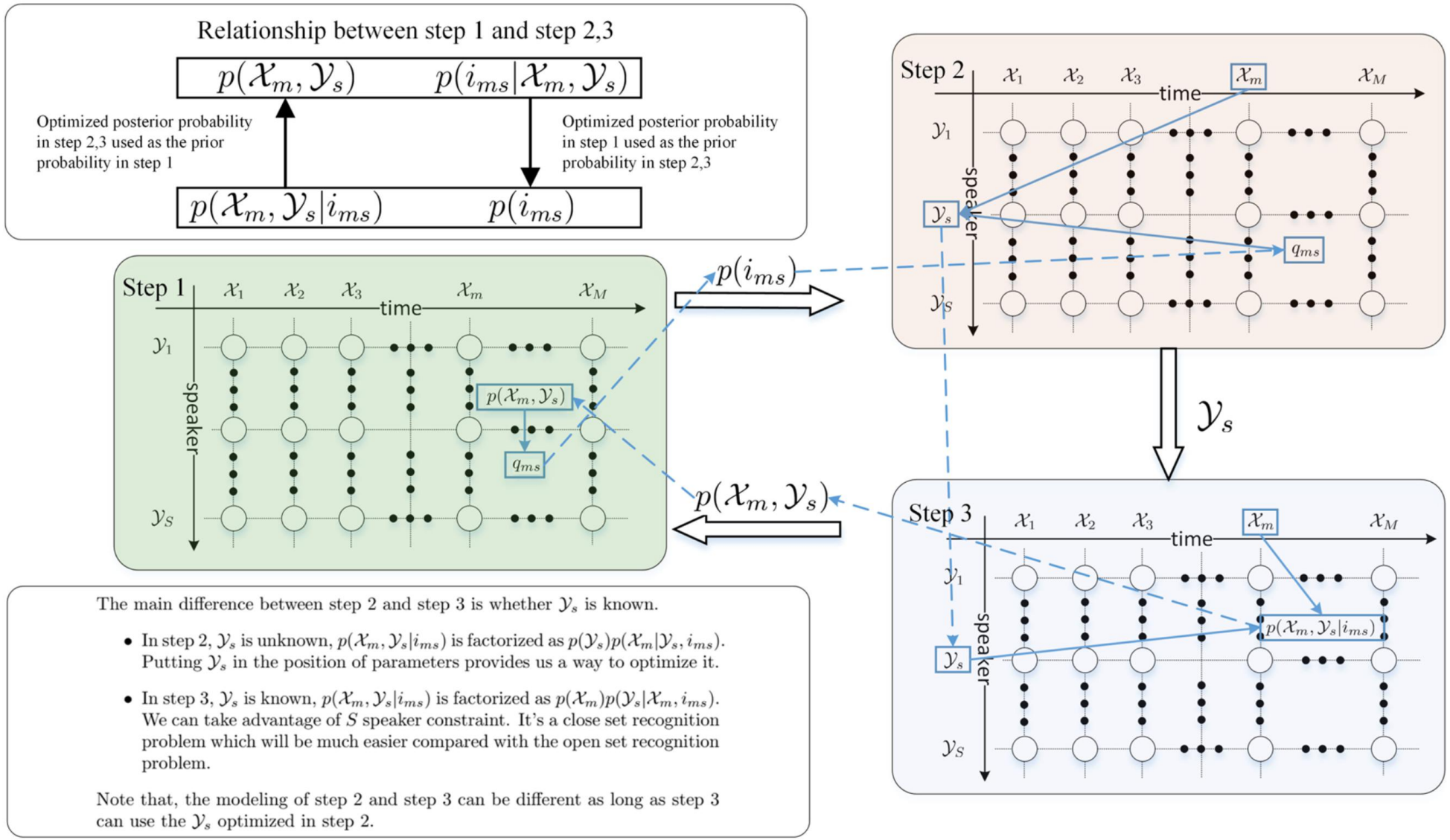
- 步骤3：

$$\max\sum_{m=1}^{M}\log\sum_{s=1}^{S}p(\mathcal{X}_m,\mathcal{Y}_s,i_{ms})$$
$$=\sum_{m=1}^{M}\log\sum_{s=1}^{S}p(i_{ms})p(\mathcal{X}_m,\mathcal{Y}_s|i_{ms})$$
$$=\sum_{m=1}^{M}\log\sum_{s=1}^{S}q_{ms}p(\mathcal{X}_m)p(\mathcal{Y}_s|\mathcal{X}_m,i_{ms})$$

给定说话人模型Y、语音段X和概率矩阵Q，估计Xm属于Ys的概率，典型的短语音说话人闭集识别任务



说话人标记



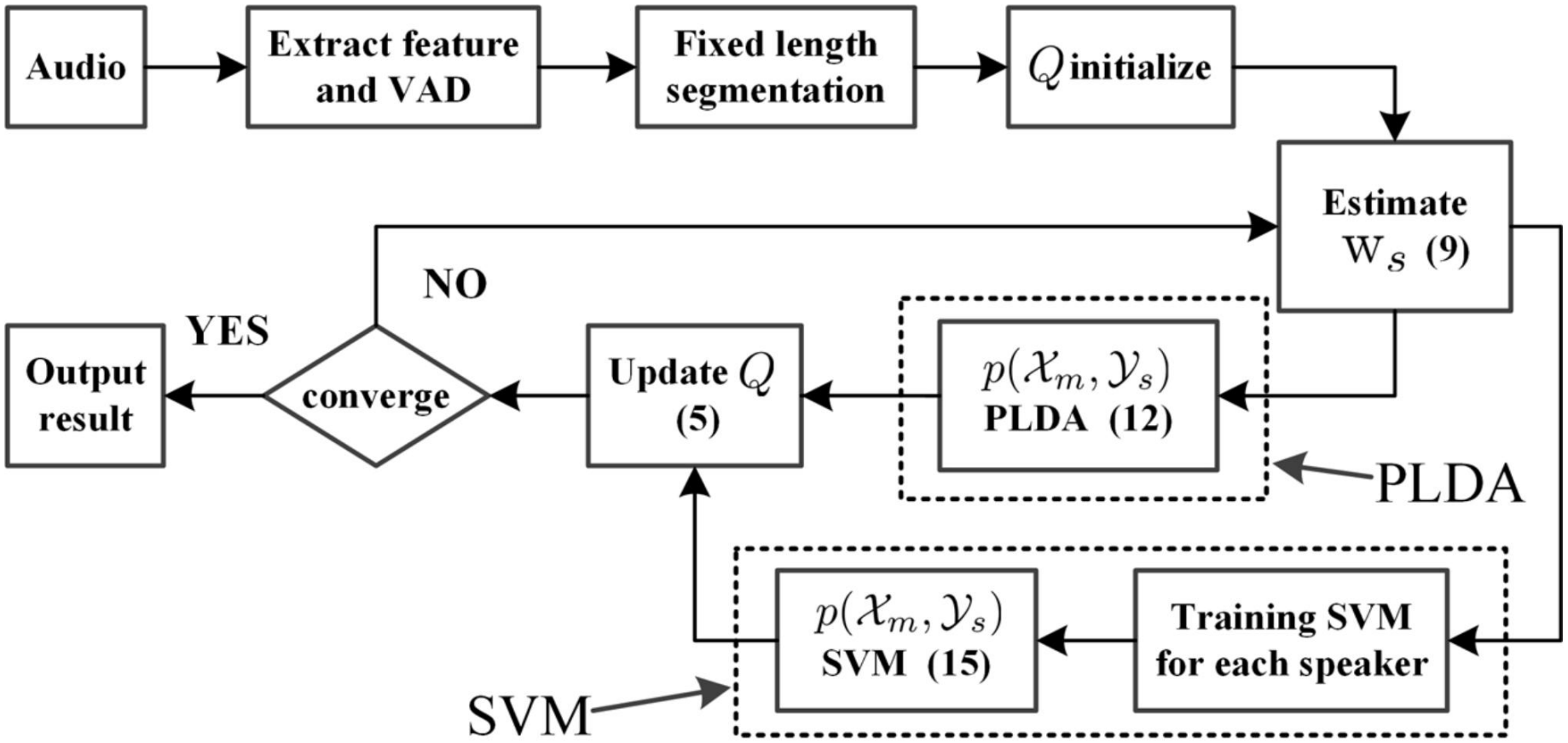
The main difference between step 2 and step 3 is whether $\mathcal{Y}_s$ is known.

- In step 2, $\mathcal{Y}_s$ is unknown, $p(\mathcal{X}_m, \mathcal{Y}_s|i_{ms})$ is factorized as $p(\mathcal{Y}_s)p(\mathcal{X}_m|\mathcal{Y}_s, i_{ms})$. Putting $\mathcal{Y}_s$ in the position of parameters provides us a way to optimize it.
- In step 3, $\mathcal{Y}_s$ is known, $p(\mathcal{X}_m, \mathcal{Y}_s|i_{ms})$ is factorized as $p(\mathcal{X}_m)p(\mathcal{Y}_s|\mathcal{X}_m, i_{ms})$. We can take advantage of S speaker constraint. It's a close set recognition problem which will be much easier compared with the open set recognition problem.

Note that, the modeling of step 2 and step 3 can be different as long as step 3 can use the $\mathcal{Y}_s$ optimized in step 2.

说话人标记

- 具体实现：混合迭代，VB-Ivec-PLDA和VB-Ivec-SVM



说话人标记

- 实验结果：NIST RT09、CallHome00

DER[%]	Speaker #	BIC	VB	LCM-Ivec		
				PLDA	SVM	Hybrid
given speaker #	-	Yes	Yes	Yes	Yes	Yes
EDI_20071128-1000	4	29.32	10.67	9.89	9.91	9.83
EDI_20071128-1500	4	35.61	48.66	19.68	19.87	17.40
IDI_20090128-1600	4	29.12	11.15	7.02	7.14	7.14
IDI_20090129-1000	4	37.27	35.85	31.99	32.37	21.82
NIST_20080201-1405	5	61.54	49.05	44.67	43.05	38.53
NIST_20080227-1501	6	40.32	39.97	24.76	25.66	13.96
NIST_20080307-0955	11	46.62	23.50	22.86	16.44	16.00
average	-	39.97	31.26	22.98	22.06	17.81

speaker #	2 (303)	3 (136)	4 (43)	5 (10)	6 (6)	7 (2)	Average
Table 2 in [54]	8.7	15.7	15.1	20.2	25.5	29.8	11.67
Figure 5 in [37] *	5.0	12.5	17.7	20.5	21.5	33.1	8.75
Table 5 in [55]	7.5	11.8	14.9	22.8	25.9	26.9	9.91
[33]	-	-	-	-	-	-	13.7
Kaldi [56]	-	-	-	-	-	-	8.69
VB+windows	6.68	14.51	18.68	26.78	24.88	25.35	10.53
LCM-Ivec-Hybrid	4.26	13.12	17.96	25.74	24.70	25.35	8.60

54. Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C.: Stream-based speaker segmentation using speaker factors and eigenvoices. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4133–4136 (2008)

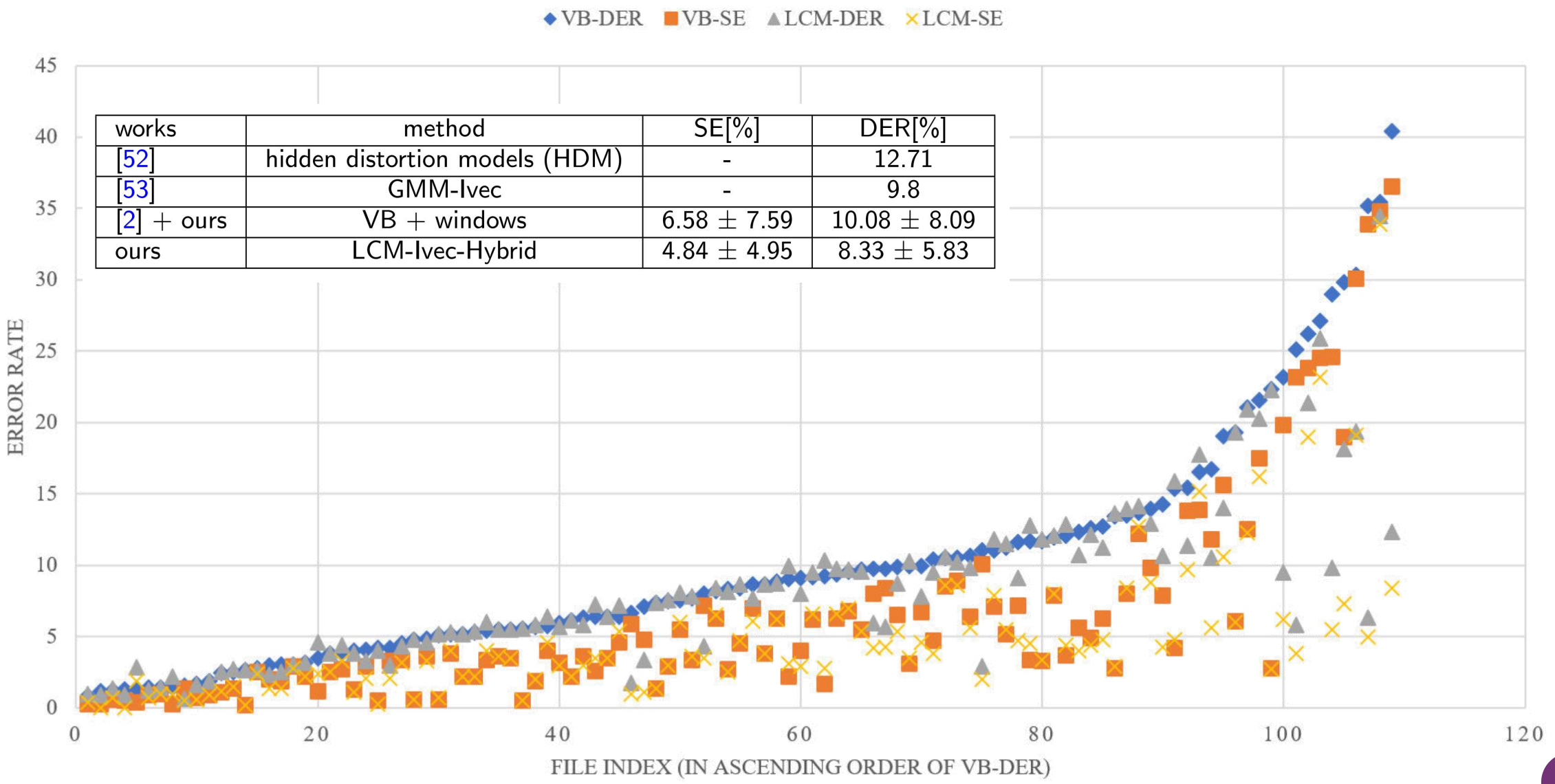
37. Shum, S.H., Dehak, N., Dehak, R., Glass, J.R.: Unsupervised methods for speaker diarization: An integrated and iterative approach. IEEE Transactions on Audio, Speech, and Language Processing **21**(10), 2015–2028 (2013). doi:10.1109/TASL.2013.2264673

55. Senoussaoui, M., Kenny, P., Stafylakis, T., Dumouchel, P.: A study of the cosine distance-based mean shift for telephone speech diarization. IEEE/ACM Transactions on Audio Speech, Language Processing **22**(1), 217–227 (2013)

33. Sell, G., Garcia-Romero, D.: Speaker diarization with plda i-vector scoring and unsupervised calibration. In: 2014

56. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011). IEEE Catalog No.: CFP11SRW-USB

说话人标记



works	method	SE[%]	DER[%]
[52]	hidden distortion models (HDM)	-	12.71
[53]	GMM-Ivec	-	9.8
[2] + ours	VB + windows	6.58 ± 7.59	10.08 ± 8.09
ours	LCM-Ivec-Hybrid	4.84 ± 4.95	8.33 ± 5.83

团队工作

- 联合学习
  - 在深度神经网络不同层面利用音素信息可提升声纹识别性能
- 对抗学习
  - 帧层面对抗学习有效
  - 模块复用与重复学习有利于强化目标信息（声纹信息）
- 说话人标记
  - 该任务近似等价于短时说话人识别任务
  - 全概率框架可融合多说话人标记结果，为避免局部最优、系统融合提供了很好的解决方法


清華大學 Tsinghua University

新疆大學 Xinjiang University

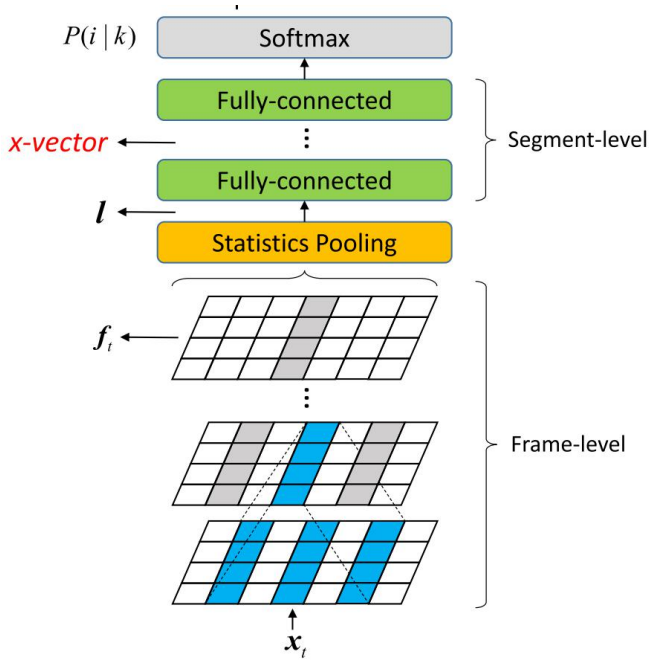
三、技术展望

超大规模说话人识别、自训练与预训练、和图神经网络

超大规模说话人识别

- 常见的声纹数据库规模

Dataset	人数	文件/人数	时长(h)
SWB	2594	10.9	1139
SRE04	310	26.4	197
SRE06	2228	15	946
SRE08	1328	20.2	724
SRE10	506	15.4	286
SRE16	221	53.5	121
SRE18	213	10.8	x
SRE19	195	11.1	x
SRE-ALL	5001	18.4	x
VOXCELEB	7146	23.2	2420
TIMIT	630	10	x
SITW	299	9.36	x



当说话人个数达到百万级以上，面临的挑战：

1、可训练问题？Softmax + Triplet Loss

2、性能是否显著下降？

1. P. Matějka *et al.*, “13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE,” *Computer Speech & Language*, vol. 63, p. 101035, Sep. 2020, doi: 10.1016/j.csl.2019.101035.

自训练与预训练

- 极少量标注数据+大量无标注数据 = SOTA声纹识别系统？

Self-training and Pre-training are Complementary for Speech Recognition

Qiantong Xu* Alexei Baevski* Tatiana Likhomanenko Paden Tomasello

Alexis Conneau Ronan Collobert Gabriel Synnaeve Michael Auli

Facebook AI Research

Abstract

Self-training and unsupervised pre-training have emerged as effective approaches to improve speech recognition systems using unlabeled data. However, it is not clear whether they learn similar patterns or if they can be effectively combined. In this paper, we show that are complementary in a vari labeled data from Libri-light achieves WERs of 3.0%/5.2 rivaling the best published sy ago. Training on all labeled

Table 1: WER on the Librispeech dev and test sets for the Libri-light low-resource labeled data setups of 10 min, 1 hour and 10 hours. As unlabeled data we use the audio of Librispeech (LS-960) or the larger LibriVox (LV-60k). ST (s2s scratch) trains a sequence to sequence model with a word-piece vocabulary on the pseudo-labeled data from random initialization while as ST (ctc ft) fine-tunes wav2vec 2.0 with the pseudo-labels using CTC and a letter-based vocabulary. All results are with language models at inference time.

Model	Unlbld data	dev clean	dev other	test clean	test other
10 min labeled					
Discr. BERT [27]	LS-960	15.7	24.1	16.3	25.2
wav2vec 2.0 [24]	LS-960	6.6	10.6	6.8	10.8
+ ST (s2s scratch)	LS-960	4.1	7.0	5.0	8.1
+ ST (ctc ft)	LS-960	3.6	6.6	4.0	7.2
wav2vec 2.0 [24]	LV-60k	5.0	8.4	5.2	8.6
+ ST (s2s scratch)	LV-60k	2.6	4.7	3.1	5.4
+ ST (ctc ft)	LV-60k	2.8	4.6	3.0	5.2
1h labeled					
Discr. BERT [27]	LS-960	8.5	16.4	9.0	17.6
wav2vec 2.0 [24]	LS-960	3.8	7.1	3.9	7.6
+ ST (s2s scratch)	LS-960	2.9	5.6	3.4	6.6
+ ST (ctc ft)	LS-960	2.8	5.5	3.1	6.3
abeled					
BERT [27]	LS-960	5.3	13.2	5.9	14.1
[4]	LS-960	23.5	25.5	24.4	26.0
vec 2.0 [24]	LS-960	2.9	5.7	3.2	6.1
(s2s scratch)	LS-960	2.5	5.1	3.5	5.9
(ctc ft)	LS-960	2.6	5.2	2.9	5.7

问题：

1、如何通过自训练和预训练，通过极少量标注数据和大量无标注数据，构建声纹识别系统

1. Q. Xu *et al.*, “Self-training and Pre-training are Complementary for Speech Recognition,” *arXiv:2010.11430 [cs, eess]*, Oct. 2020, Accessed: Nov. 20, 2020. [Online]. Available: http://arxiv.org/abs/2010.11430.

图神经网络

- 现状：判断两段语音是否属于同一个说话人，Cosine/Inner Product/PLDA
- 未来：从网络视角看待声纹识别

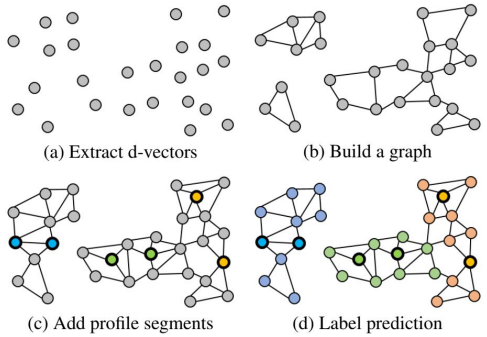


Figure 1: *Overview of the proposed method: (a) extract d-vectors of audio segments with a pre-trained speaker embedding model; (b, c) build a graph of audio segments based on pair-wise similarities of the corresponding d-vectors, using both profile and test audio segments; (d) predict labels for test audio segments by graph-based semi-supervised learning methods.*

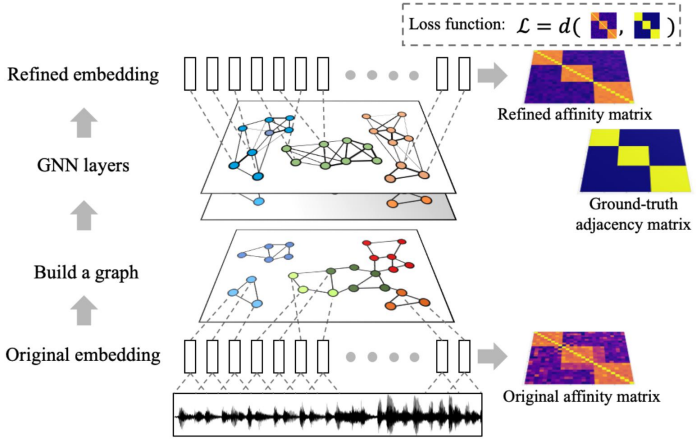


Table 1. DER (%) on the NIST SRE 2000 CALLHOME dataset. SC refers to spectral clustering and AHC to agglomerative hierarchical clustering.

	Method	DER(%)
Baseline	x-vector + PLDA + AHC *(5-fold)*	8.64
	x-vector + PLDA + SC *(5-fold)*	8.05
Recent Work	Wang *et al.* [12]	12.0
	Sell *et al.* [33]	11.5
	Romero *et al.* [34]	9.9
	Zhang *et al.* [13] *(5-fold)*	8.5
	Lin *et al.* [26] *(5-fold)*	7.73
Ours	GNN based *(5-fold)*	**7.24**

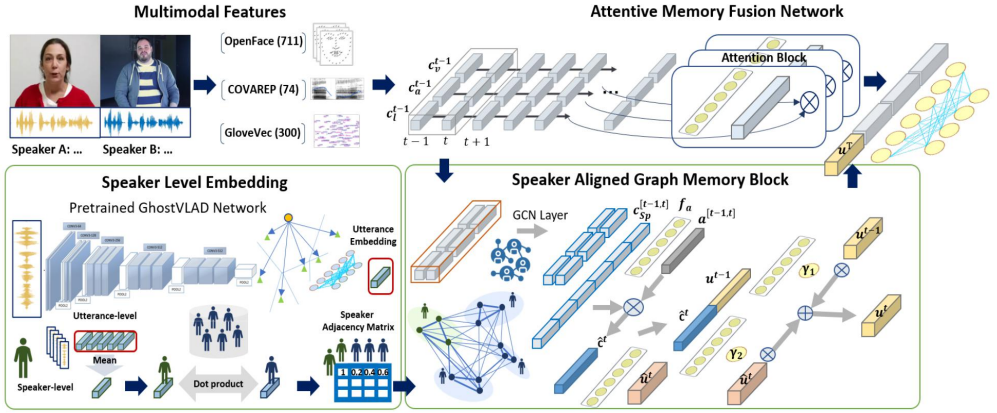


Figure 1: *Our framework SaGMN has a multimodal backbone network with a speaker aligned memory block. The similarity of speaker embeddings extracted from pre-trained speaker recognition network are used to derive adjacency matrix for the graph convolutional layer in the memory block. The resulting memory vector and multimodal attended vectors are used for the final emotion recognition.*

1. J. Wang, X. Xiao, J. Wu, R. Ramamurthy, F. Rudzicz, and M. Brudno, "Speaker Diarization with Session-Level Speaker Embedding Refinement Using Graph Neural Networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7109–7113.
2. J. Wang, X. Xiao, J. Wu, R. Ramamurthy, F. Rudzicz, and M. Brudno, "Speaker Attribution with Voice Profiles by Graph-Based Semi-Supervised Learning," in *Interspeech 2020*, Oct. 2020, pp. 289–293.
3. J.-L. Li and C.-C. Lee, "Using Speaker-Aligned Graph Memory Block in Multimodally Attentive Emotion Recognition Network," in *Interspeech 2020*, Oct. 2020, pp. 389–393.

清華大學 Tsinghua University

新疆大學 Xinjiang University

谢 谢！