声纹和语种识别中的多任务学习研究

李 琳

厦门大学智能语音实验室
http://speech.xmu.edu.cn

2020.11

纲要

- 研究背景
- 相关研究
- 学习机制
- 思考与分析
  - 声纹识别
  - 语种识别
- 总结与展望

研究背景——为什么引入“多任务”

多任务学习

- **相关任务**
- **共享信息**

归纳偏置
an inductive bias

改善性能
More generative performance

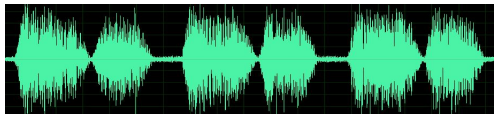
Caruana R (1998), Multitask learning. In: Learning to learn, pp 95–133. Springer.

研究背景——如何引入“多任务”

厦門大學 XIAMEN UNIVERSITY

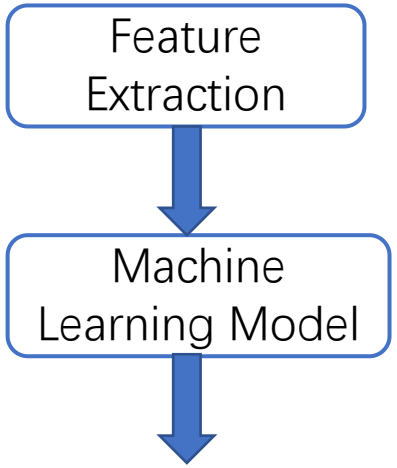
- 人类能力
  - 触类旁通
  - 举一反三
  - 问牛知马
  - …

- 语音信号



  - 文本信息
  - 说话人身份
  - 语种
  - 情感
  - 性别
  - ……

- 语音相关任务



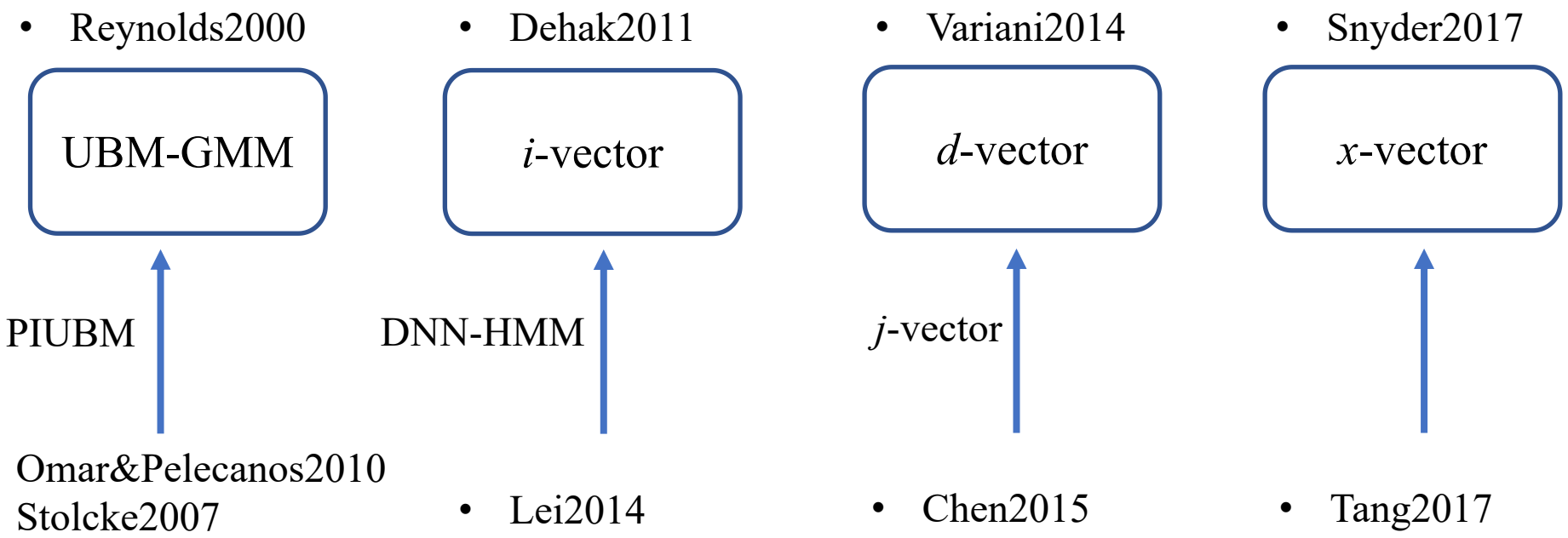
  - 语音合成
  - 语音识别
  - 声纹识别
  - 语种识别
  - 情感识别
  - 性别识别
  - ……

多任务学习

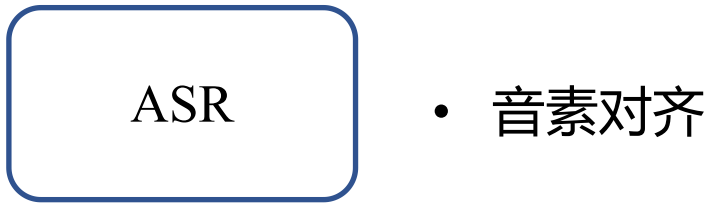
相关任务
共享信息

相关研究——声纹识别中的“多任务”

- 经典的声纹识别框架

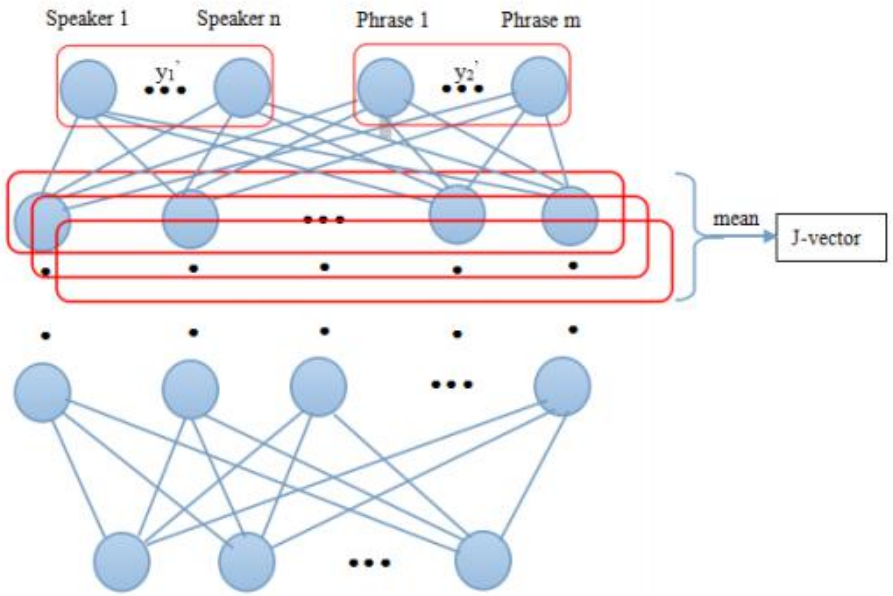


- 音素相关信息的引入



相关研究——声纹识别中的“多任务”

- d-vector with phonetic-aware information[Chen2015]

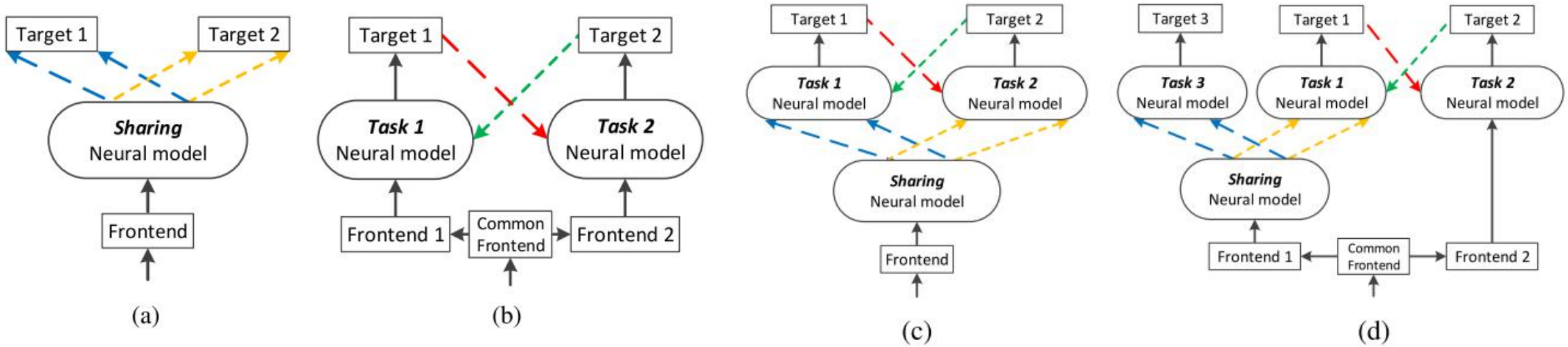


- 多任务+深度学习 任务设定位： 获取说话人特征与本文信息
- 最后一个隐藏层提取 *j-vector*

Nanxin Chen, Yanmin Qian and Kai Yu, "Multi-Task Learning for Text-dependent Speaker ", *INTERSPEECH2015*.

相关研究——声纹识别中的“多任务”

- Collaborative learning with phonetic-aware information[Tang2017]

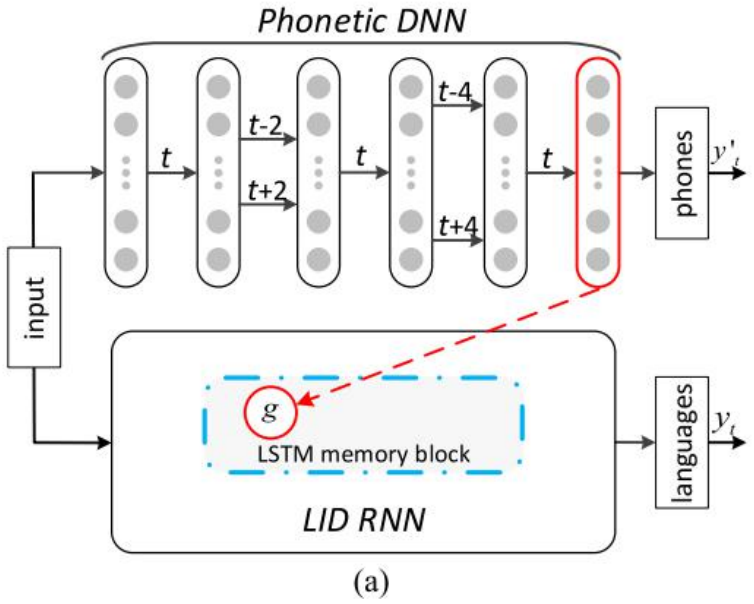


- 分析语音识别和声纹识别多任务学习的协同模式

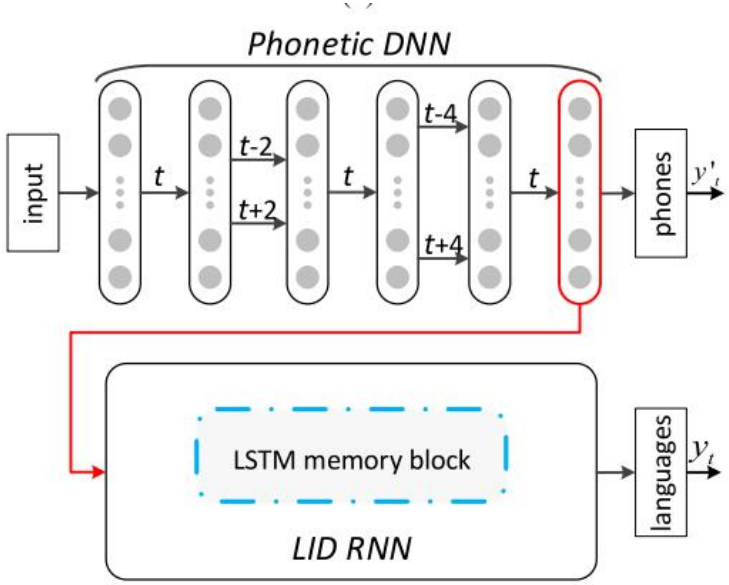
Zhiyuan Tang, Lantian Li, Dong Wang, and Ravichander Vipperla, “ Collaborative Joint Training With Multitask Recurrent Model for Speech and Speaker Recognition”, *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2017.*

相关研究——语种识别中的“多任务”

- Phonetic Temporal Neural Model for Language Identification[Tang2018]



(a)



(b)

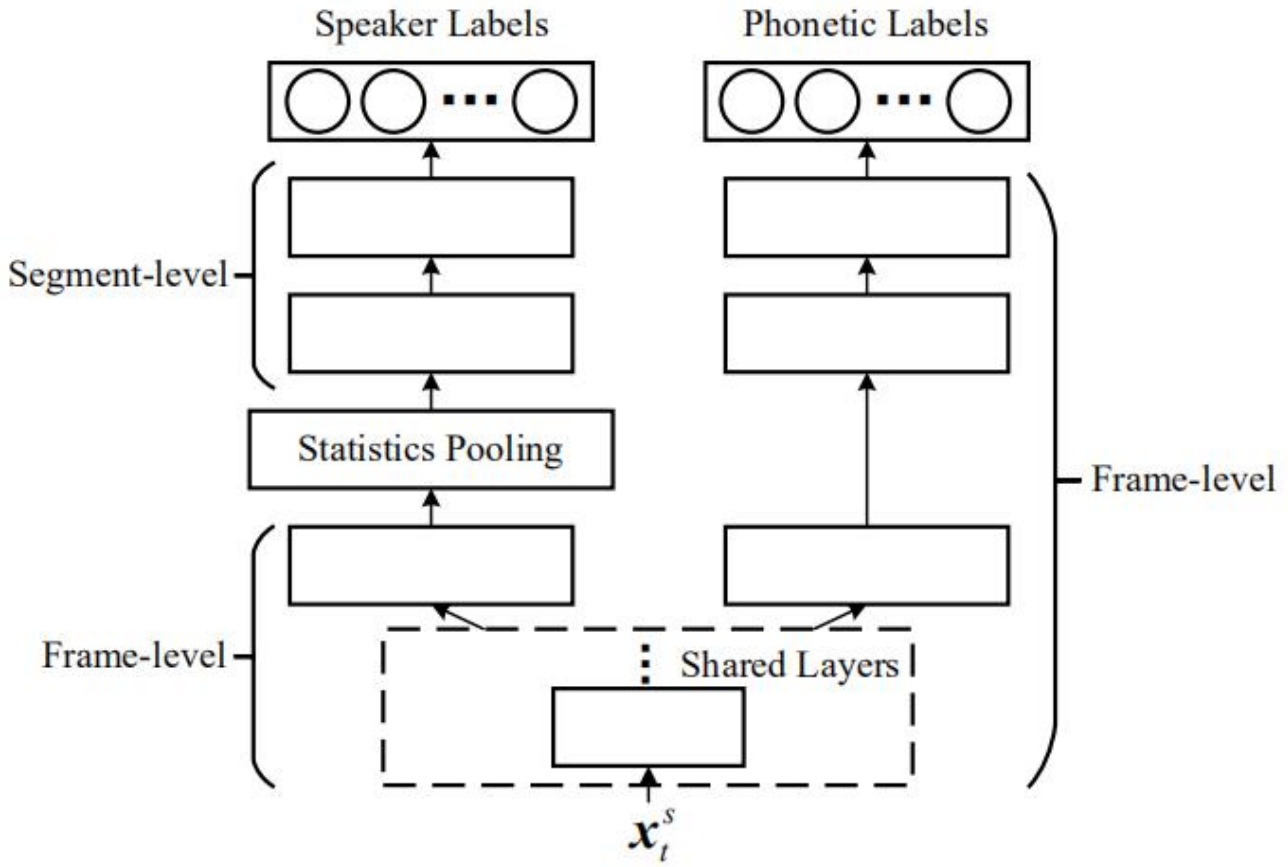
- 基于DNN的音素分类网络提供音素特征
- 基于RNN的语种识别网络（引入音素特征）

Zhiyuan Tang, Dong Wang, Yixiang Chen, Lantian Li, and Andrew Abel, “Phonetic Temporal Neural Model for Language Identification”, *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018.*

相关研究——声纹识别中的“多任务”

- X-vector with phonetic-aware information[Liu2018]

- 文本相关声纹识别
- 相关任务：
  - 说话人识别+帧级别音素识别
- 共享信息
  - 帧级别TDNN权重共享



Yi Liu , Liang He , Jia Liu, and Michael T. Johnson, "Speaker Embedding Extraction with Phonetic ", *INTERSPEECH2018*.

相关研究——声纹识别中的“多任务”

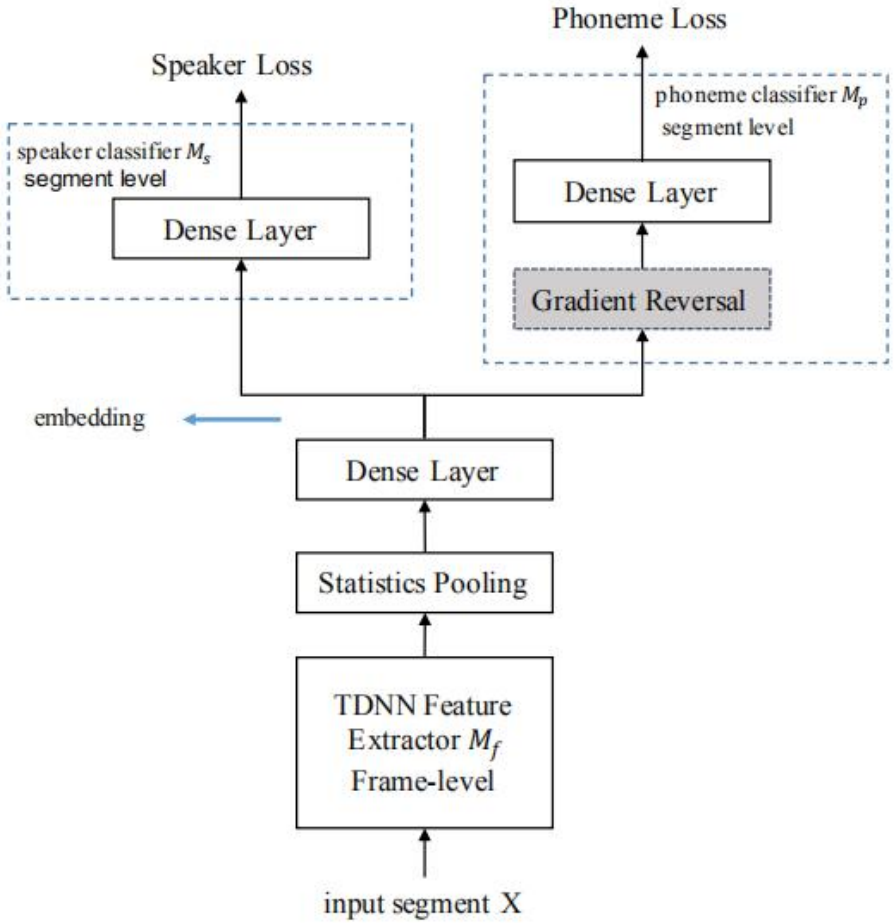
- Adversarial training with phonetic-aware information[Wang2019]

- 文本无关声纹识别
- 段级别音素标签

$$\mathbf{y}^p = \{y_1, y_2, \ldots, y_C\}$$
$$y_c = \frac{N_c}{N}$$

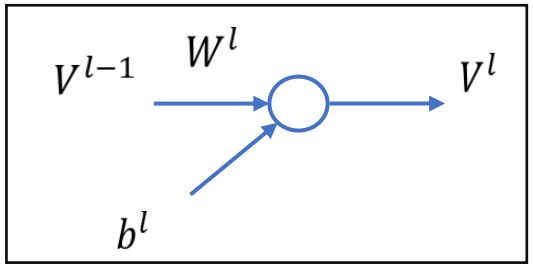
- 梯度反转层（Gradient Reversal Layer, GRL）
  - 段级别：抑制音素信息

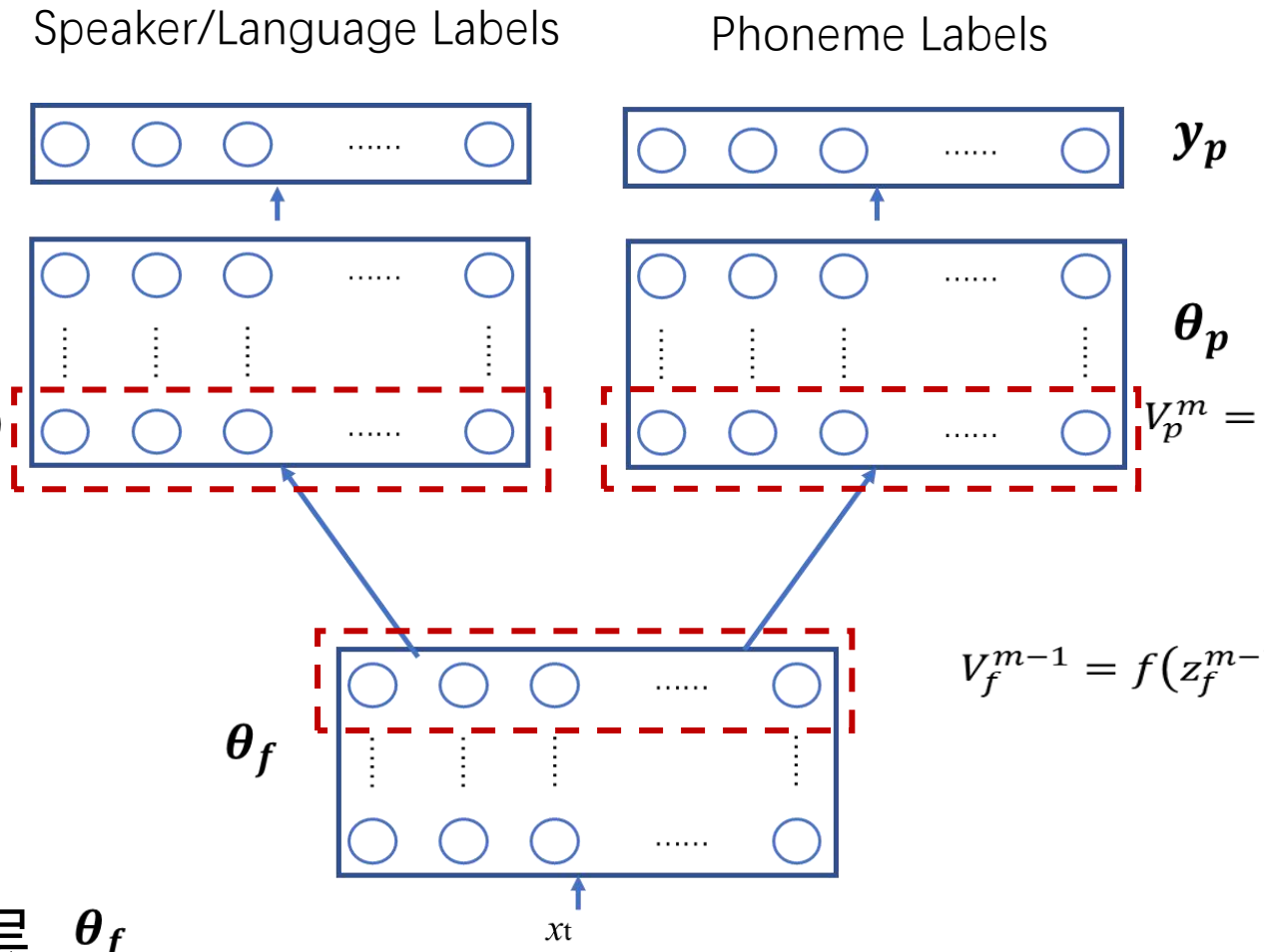
- 组合式：帧级别多任务+段级别抑制



Shuai Wang, Johan Rohdin, Lukas Burget, Oldrich Plchot, Yanmin Qian, Kai Yu, Jan Honza Cernock, "On the Usage of Phonetic Information for Text-independent Embedding Extraction", *INTERSPEECH2019*.

多任务学习机制-前向传播





$$V_c^m = f(z_c^m) = f(W_c^m V_f^{m-1} + b_c^m)$$

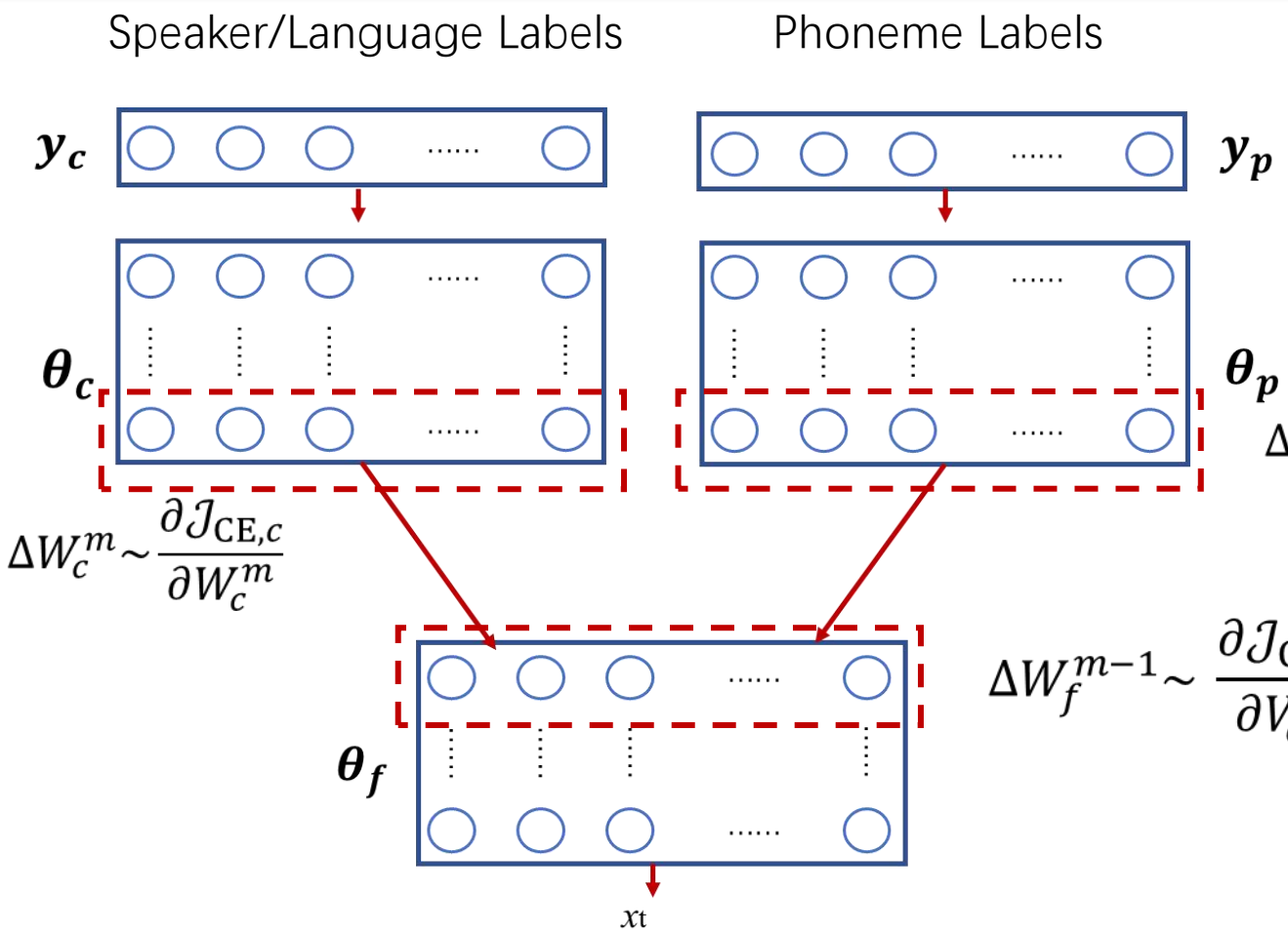
$$V_p^m = f(z_p^m) = f(W_p^m V_f^{m-1} + b_p^m)$$

$$V_f^{m-1} = f(z_f^{m-1}) = f(W_f^{m-1} V_f^{m-2} + b_f^{m-1})$$

共享权重的网络层 θ_f

任务相关的网络层 θ_c θ_p

多任务学习机制——反向传播引入音素作用



$$\mathcal{J} = \mathcal{J}_{\mathrm{CE},c} + \alpha \times \mathcal{J}_{\mathrm{CE},p}$$

$$\Delta W_f^{m-1} \sim \frac{\partial \mathcal{J}_{\mathrm{CE},c}}{\partial V_c^m}\frac{\partial V_c^m}{\partial V_f^{m-1}}\frac{\partial V_f^{m-1}}{\partial W_f^{m-1}} \pm \alpha \times \frac{\partial \mathcal{J}_{\mathrm{CE},p}}{\partial V_p^m}\frac{\partial V_p^m}{\partial V_f^{m-1}}\frac{\partial V_f^{m-1}}{\partial W_f^{m-1}}$$

共享权重的网络层引入全监督/半监督权重学习

思考与分析

- 多任务学习的优势
  - 音素HMM的概率密度函数(*pdf*)
  - 帧级别音素对齐信息
  - 鲁棒的深度特征表征
    - 紧凑——类内距离小 S_w
    - 可分离——类间距离大 S_b

思考与分析——文本无关声纹识别[Wang2019]

表1 Voxceleb1测试集上的实验结果（Voxceleb1&2训练）

System	EER%	minDCF(0.01)	minDCF(0.1)
TDNN x-vector baseline	3.73	0.389	0.192
FRM-MT	3.38	0.357	0.180
FRM-GRL	5.24	0.502	0.269
SEG-MT	3.71	0.327	0.175
SEG-GRL	3.35	0.332	0.159
COMBINE (FRM-MT&SEG-GRL)	3.17	0.336	0.163

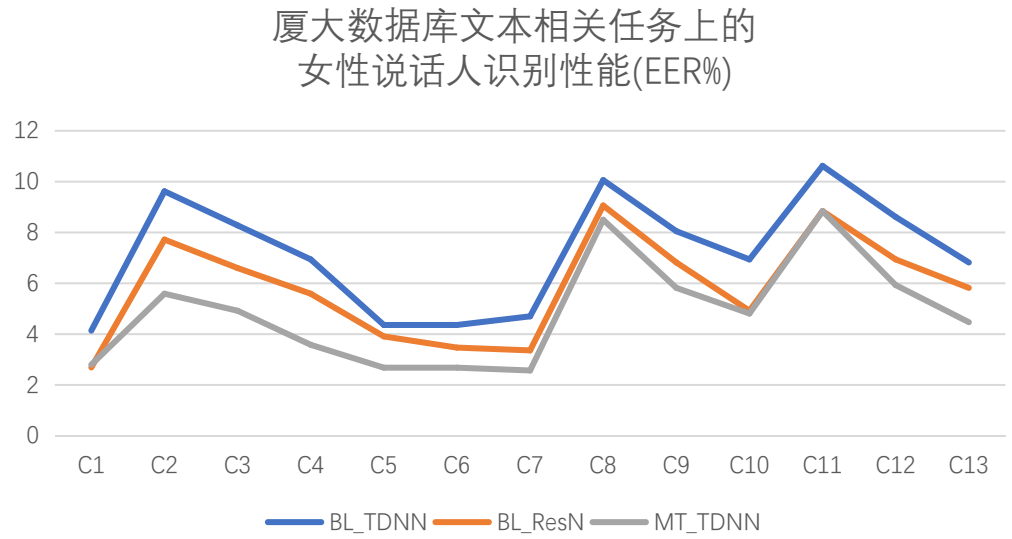
- FRM-MT v.s.FRM-GRL：引入帧级别的音素对齐信息是有益的
- FRM-MT v.s. SEG-MT：帧级别的音素对齐信息更重要
- FRM-MT v.s. SEG-GRL:强调帧级别音素对齐与抑制全局音素信息具有近似的效果
- COMBINE: 音素双重作用——帧级别上的音素对齐与全局抑制

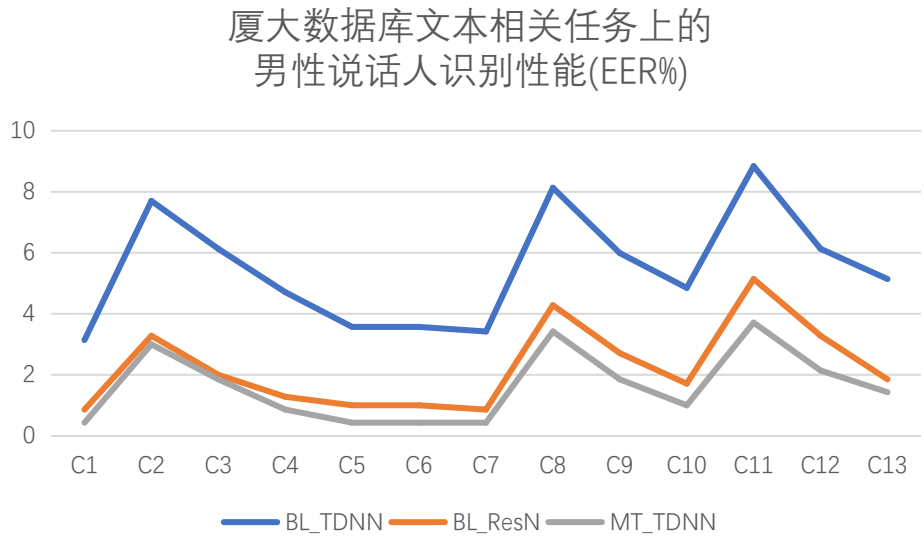
Shuai Wang, Johan Rohdin, Lukas Burget, Oldrich Plchot, Yanmin Qian, Kai Yu, Jan Honza Cernock, "On the Usage of Phonetic Information for Text-independent Embedding Extraction", *INTERSPEECH2019*.

思考与分析——文本相关声纹识别

厦門大學 XIAMEN UNIVERSITY

- 在厦大数据库文本相关任务上，不同性别不同噪声情况下采用多任务学习方法，EER值都大幅度降低。





- 训练集769人，测试集59人
- C1~C13代表13种不同噪声条件
- BL_TDNN代表TDNN基线系统，BL-ResN代表ResNet基线系统，MT-TDNN代表多任务学习TDNN系统。

思考与分析——语种识别中的“多任务”

厦門大學 XIAMEN UNIVERSITY

- 东方语种识别竞赛(OLR Challenge 2018)
  - Task1（短语音语种识别）第一名
  - Task2（混淆语言语种识别）第二名
  - Task3（开集语种识别）第二名
- 东方语种识别竞赛(OLR Challenge 2019)
  - Task1（短语音语种识别）第三名
  - Task2（跨信道语种识别）第三名
  - Task3（零资源语种识别）第一名
- 东方语种识别竞赛(OLR Challenge 2020)
  - Task1（跨信道语种识别）
  - Task2（开集方言识别）
  - Task3（带噪语种识别）

Task 1: short-utterance
Ranking 1 -- 10 (Results)

Ranking	Team Name	Cavg	EER%
1	xmuspeech	0.0462	4.59
2	DKU-Tencent	0.0499	5.01
3	DKU-Tencent-Babel	0.0512	5.19
4	NetEase AI-Speech Group	0.0548	5.53
5	DKU-Tencent-Owl	0.0583	6.06
6	SASI	0.0631	6.46
7	DoubleYuan	0.0701	7.12
8	SAIT	0.0885	8.72
9	Innovem Tech	0.0971	10.04
10	anonymous	0.1153	11.59

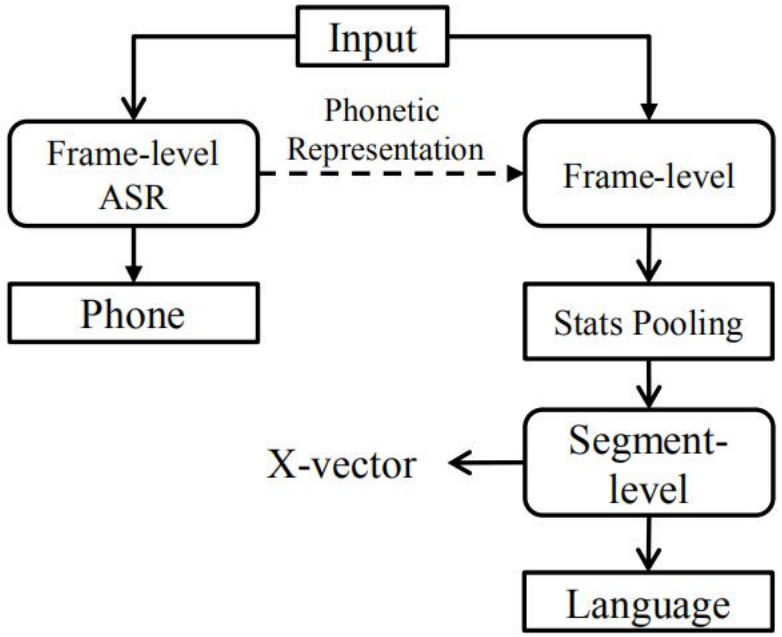
Task 3: zero-resource
Top 5

Ranking	Team Name	Institute	Participants	Cavg	EER%
1	xmuspeech	厦门大学	李铮，赵淼，李静，邵艺铭，李琳	0.0113	1.13
2	Royal Flush	浙江核新同花顺网络信息股份有限公司	胡新辉，王鼎	0.0777	7.57
3	Paic-LiangpiXishi	Ping An Technology (Shenzhen) Company Limited	Ruizhang Wang、Yangli Wang、Chong Qin、Yayun Zhou	0.1098	10.7
4	Siplab-IITH	Indian Institute of Technology Hyderabad	Shaik Mohammad Rafi,Sri Rama Murty Kodukula,Gundluru Ramesh	0.1837	18.98
5	madeinchina	中国传媒大学	周晓星、冯芝金、神瑞雪	0.2129	21.5

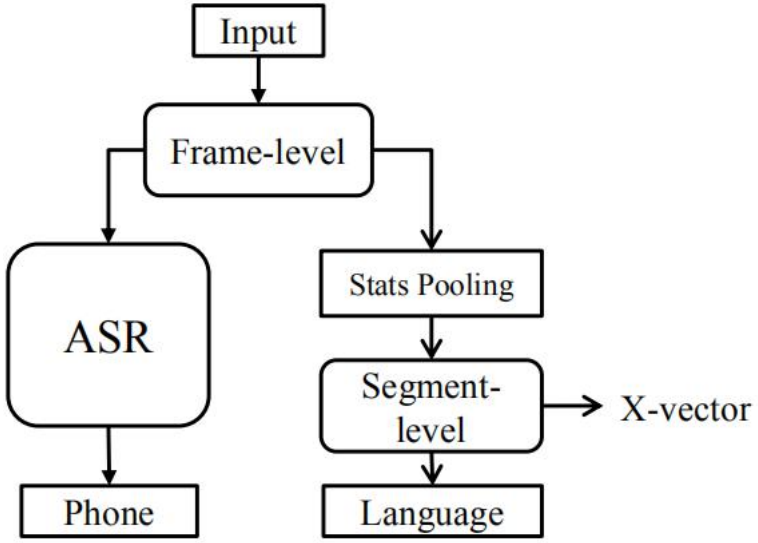
思考与分析——语种识别中的“多任务”

厦門大學 XIAMEN UNIVERSITY

- 不同的语言种类存在不同的音素分布特点与音素时序关系



(a) with the phonetic representation from ASR network(PN)



(b) multi-task learning for the language classification(MTL)

Miao Zhao , Rongjin Li , Shijiang Yan , Zheng Li , Hao Lu , Shipeng Xia , Qingyang Hong, and Lin Li, "Phone-Aware Multi-task Learning and Length Expanding for Short-Duration Language", *APSIPA2019*.

思考与分析——语种识别中的“多任务”

厦門大學 XIAMEN UNIVERSITY

- Oriental Language Recognition
  - 10种东方语种，有词典但缺少音素信息
  - 官方提供THCH30中文数据库
  - 采用THCH30训练ASR
  - 使用GMM-HMM获得帧级别对齐信息
  - 后端采用LR分类器

表3 OLR2018短语音测试集性能EER%

Datasets	Baseline	PN	MTL
Eval (AP18-OLR-TEST-Task1)	6.92	6.62	**6.50**
Dev-all (AP17-OLR-TEST)	1.72	1.64	**1.50**

Miao Zhao , Rongjin Li , Shijiang Yan , Zheng Li , Hao Lu , Shipeng Xia , Qingyang Hong, and Lin Li, "Phone-Aware Multi-task Learning and Length Expanding for Short-Duration Language", *APSIPA2019*.

思考与分析——语种识别中的“多任务”

表4 OLR多任务学习中实验数据一览表

	Task	Laguages	Uttrances	Hours	Channel
Training Set (AP16-OL7, AP17-OL3)		Mandarin, Cantonese, Indonesian, Japanese,Russian, Korean, Vietnamese,Kazakh, Tibetan, and Uyghur	282,855	140.2	Mobile
Test Set	Short Utterance (AP18-Test)	Mandarin, Cantonese, Indonesian, Japanese,Russian, Korean, Vietnamese,Kazakh, Tibetan, and Uyghur	21,456	5.8	Mobile
	Cross-channel (AP19-Test)	Cantonese, Indonesian, Japanese, Russian, Korean, and Vietnamese	10,800	8.9	Unknown

评价指标：Cavg，EER

$$C_{avg} = \frac{1}{N}\sum_{L_t}\left\{P_{Target}P_{Miss}(L_t) + \sum_{L_n}P_{Nontarget}P_{FA}(L_t, L_n)\right\}$$

思考与分析——语种识别中的“多任务”

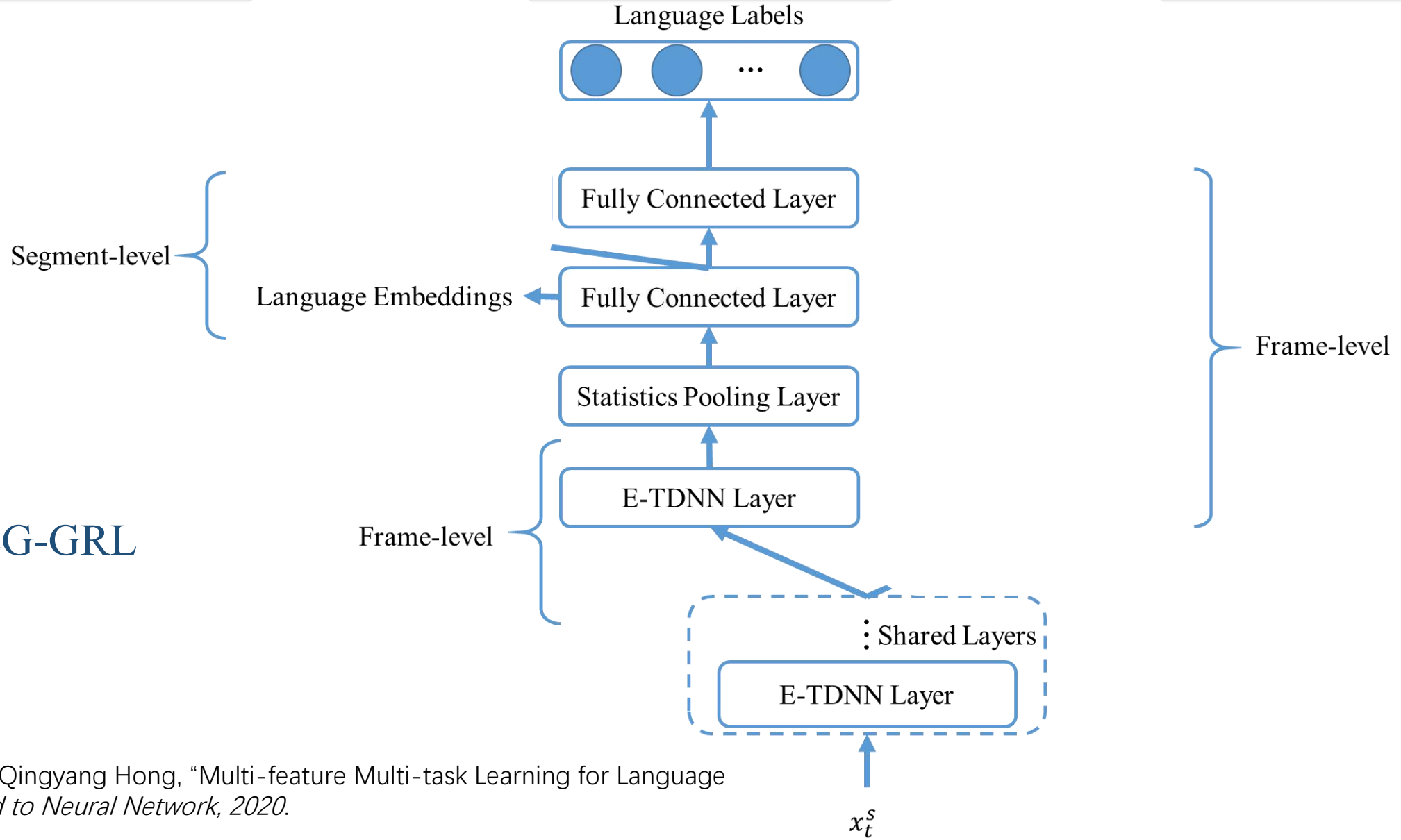
表5 语种识别多任务学习中采用不同数据库训练ASR的性能分析

System	Short Utterance		Cross-channel	
	Cavg	EER%	Cavg	EER%
TOP1 Primary Sys.	0.0462	4.59	0.2008	20.24
TOP2 Primary Sys.	0.0499	5.01	0.2713	27.69
TOP3 Primary Sys.	0.0512	5.19	0.2741	27.44
E-TDNN x-vector baseline	0.0540	5.64	0.2694	26.94
FRM-MT (THCHS30)	0.0411	4.19	0.1785	18.07
FRM-MT (Librispeech)	0.0411	4.19	0.1633	16.57

- THCHS30中文数据库训练产生 3464 个多音子HMM观测值*pdf* 标签
- Librispeech英文数据库训练产生 5704 个多音子HMM观测值*pdf* 标签

思考与分析——语种识别中的“多任务”

- FRM-MT
- SEG-MT
- SEG-GRL
- FRM-MT&SEG-GRL



Lin Li, Zheng Li, Yan Liu, Qingyang Hong, “Multi-feature Multi-task Learning for Language Recognition”, *Submitted to Neural Network, 2020*.

思考与分析——语种识别中的“多任务”

厦門大學 XIAMEN UNIVERSITY

- FRM-MT：
  - 引入帧级别的音素对齐信息是有益的
- FRM-MT v.s. SEG-MT:
  - 帧级别的音素对齐信息更重要
- FRM-MT v.s. SEG-GRL:
  - 强调帧级别音素对齐与抑制全局音素信息具有近似的效果
- COMBINE:
  - 音素双重作用——帧级别的音素对齐与全局抑制

表6 语种识别测试集上的实验结果

System	Short Utterance		Cross-channel	
	Cavg	EER%	Cavg	EER%
E-TDNN x-vector baseline	0.0540	5.64	0.2694	26.94
FRM-MT	0.0411	4.19	0.1633	16.57
SEG-MT	0.0426	4.42	0.2148	21.67
SEG-GRL	0.0418	4.24	0.2007	20.16
COMBINE (FRM-MT&SEG-GRL)	**0.0380**	**3.81**	**0.1517**	**15.04**

-采用Librispeech训练ASR音素对齐信息

思考与分析——语种识别中的“多任务”

- 类内距离
  - FRM-MT更加紧凑
  - COMBINE的最小
- 类间距离
  - 整体分布空间变化
  - 伪音素标签的影响
- 采用LR后端处理
  - 可只关注类内距离

表6 语种识别测试集embedding聚类情况

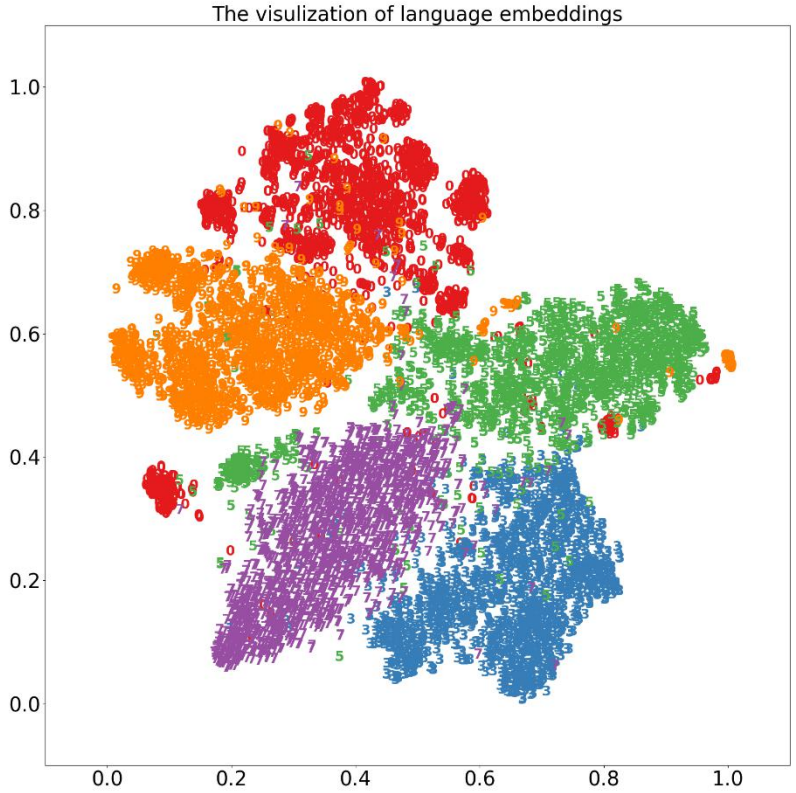
System	Short Utterance		Cross-channel	
	Sw	Sb	Sw	Sb
E-TDNN x-vector baseline	10.867	119.124	11.313	324.398
FRM-MT	4.995	24.687	5.319	72.034
SEG-MT	6.634	47.533	6.408	131.601
SEG-GRL	6.288	47.21	6.064	122.626
COMBINE (FRM-MT&SEG-GRL)	**1.862**	10.024	**1.808**	28.213

-采用Librispeech训练ASR音素对齐信息

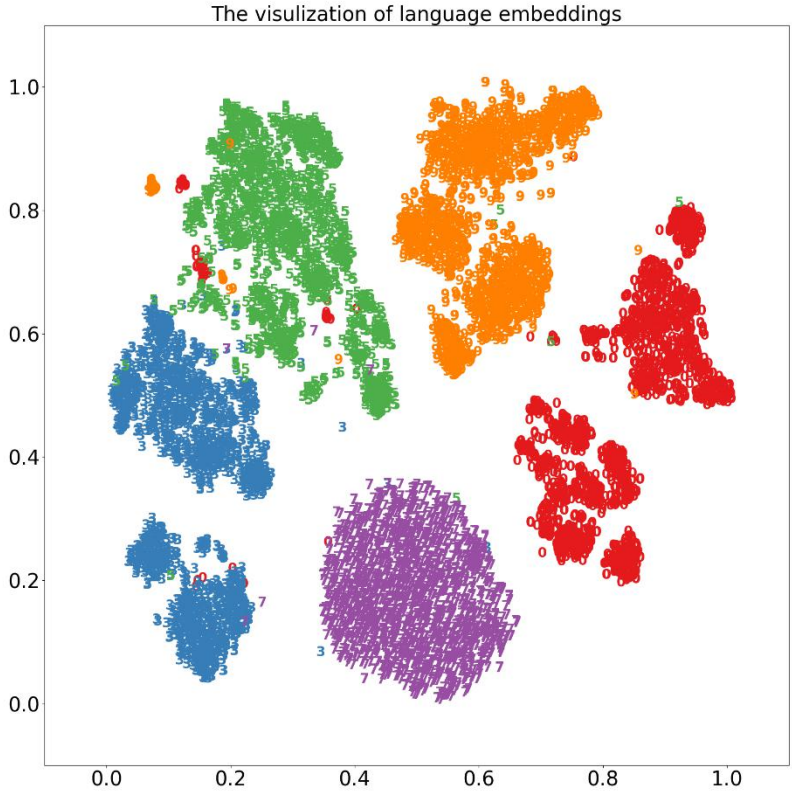
思考与分析——语种识别中的“多任务”

厦門大學 XIAMEN UNIVERSITY

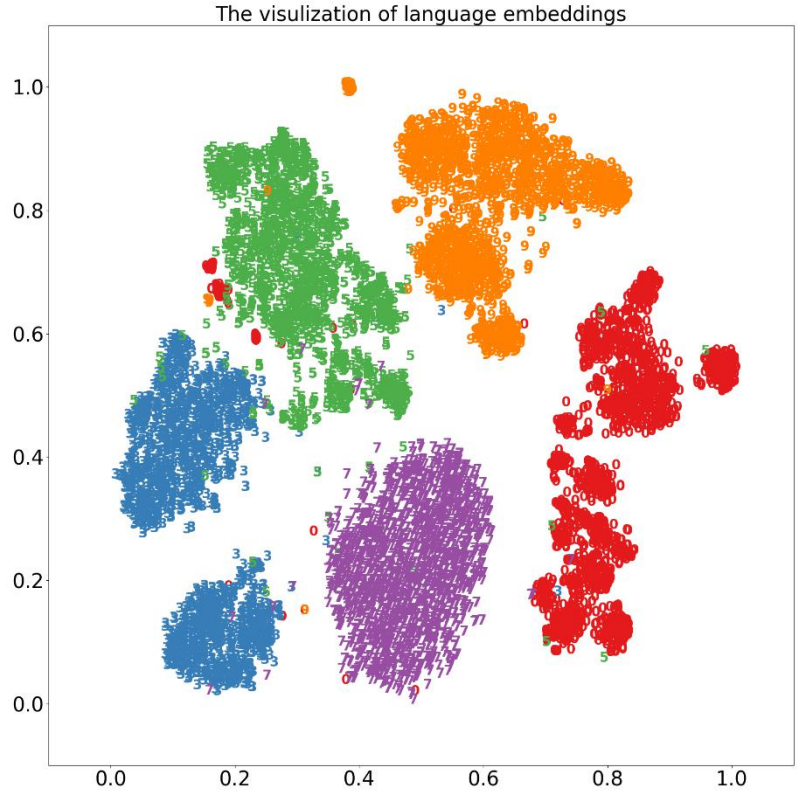
- Cross-channel情况下语种embedding的t-SNE图



(a) E-TDNN x-vector baseline



(b) FRM-MT



(c) FRM-MT&SEG-GRL

总结与展望

- 音素分类网络带来每帧的重要性判断
  - 优势：无监督->有监督
  - 局限：ASR对齐精度影响
- 声纹识别/语种识别引入多任务学习机制可改善识别性能
  - FRM-MT/SEG-MT/SEG-GRL/COMBINE
  - GRL实现段级别对抗
- 多任务学习改善深度特征的类别区分能力

总结与展望

- 展望
  - 多任务学习权重的自适应优化
  - 低资源语种任务中端到端ASR模型的训练与对齐
  - 音素信息与pooling层的结合
  - ……

THANKS!

敬请指正

厦门大学智能语音实验室
speech.xmu.edu.cn