语言资源高精尖创新中心
Beijing Advanced Innovation Center for Language Resources

基于发音属性特征和时延神经网络的语种识别研究

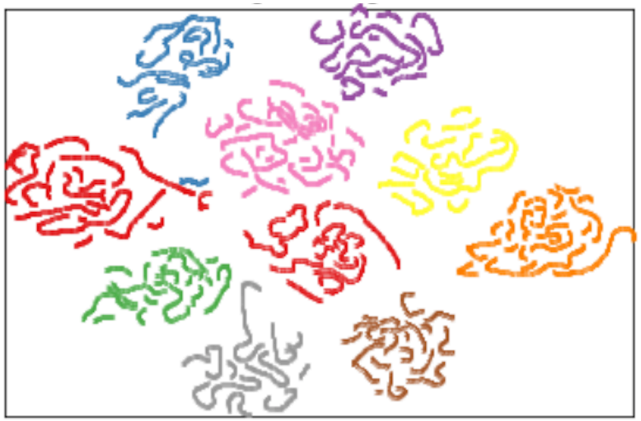
报告人：于嘉威
智能语音习得实验室（SAIT）
北京语言大学
2019.08.16

总述

- 引言
- 本研究所提出的语种识别系统
- 实验设计
- 实验结果
- 结论

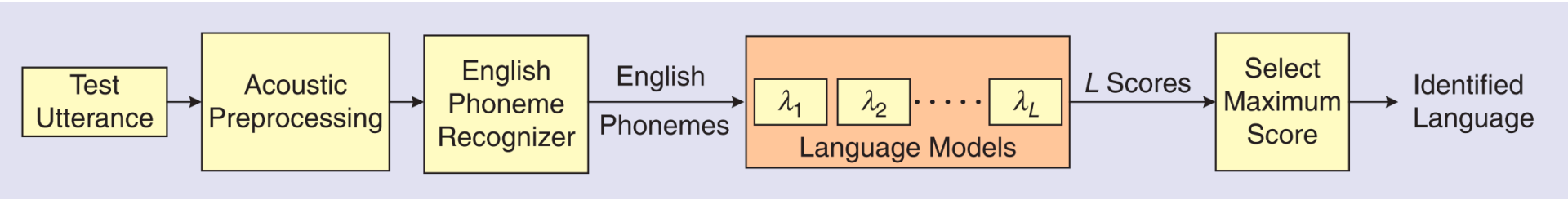
语种识别方法的分类

- 基于频谱（spectrum）



- 基于标识（token）

基于标识的语种识别方法



(*Eliathamby Ambikairajah et al., 2011*)

- 音素（phone）
  - PRLM、PPRLM (*M. A. Zissman et al., 1996*)
- 发音属性特征（Articulatory Features, AFs）
  - (*Sabato Marco Siniscalchi et al., 2009*)

基于标识方法存在的问题

- 系统性能依赖标识识别器的准确率 (*Haizhou Li et al., 2013*)
- n-gram语言模型建模标识，数据稀疏问题 (*Sabato Marco Siniscalchi et al., 2009*)

本研究所提出的语种识别系统

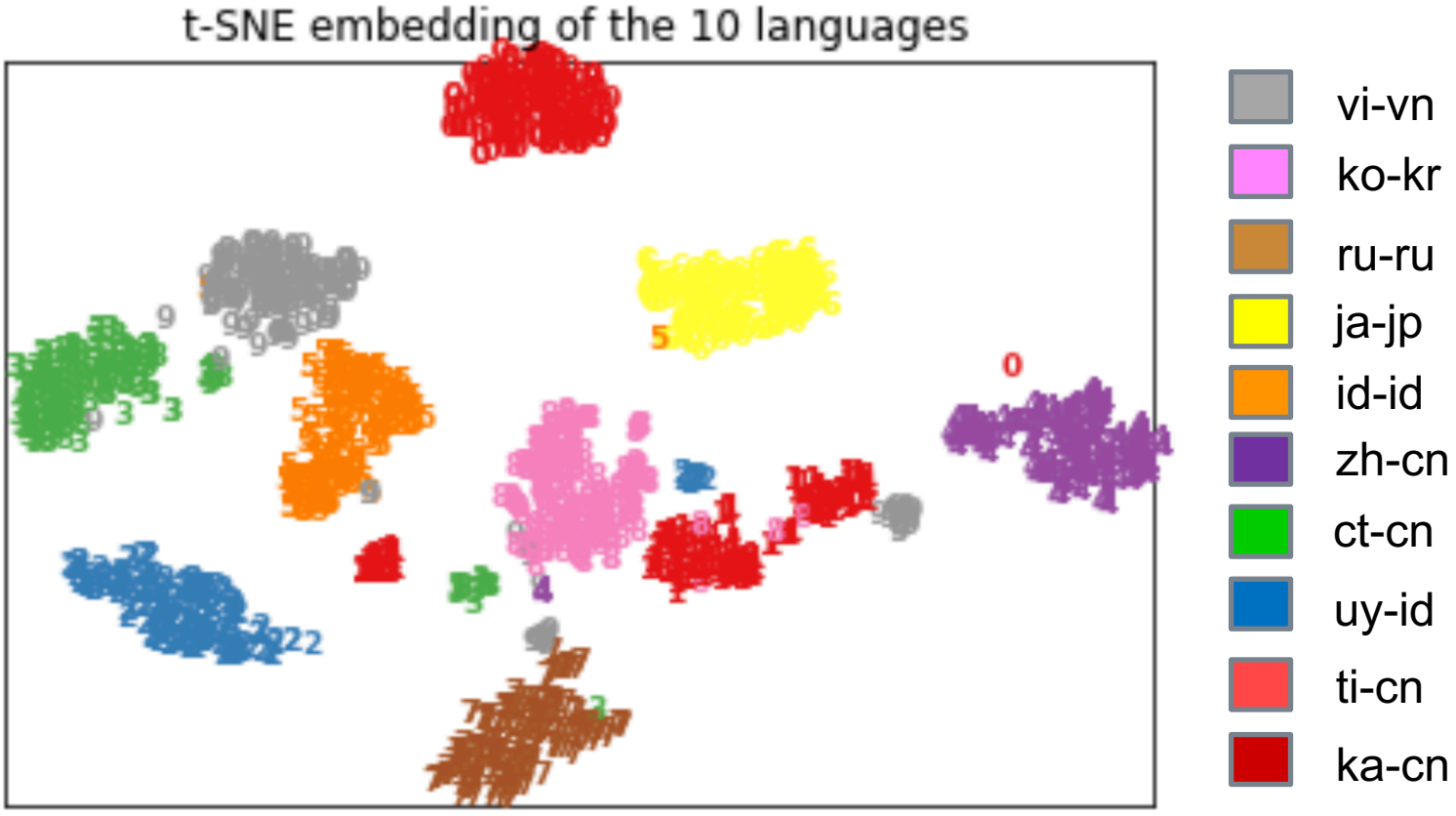
- 发音属性特征（Articulatory Features, AFs）
- 时延神经网络（Time Delay Neural Network，TDNN）

发音属性特征

- 定义：发声器官在发某个特定的音素时所引发的声道的变化。
  - 特点：跨语言（低资源）、颗粒度更小
  - 发音属性的识别准确率高于音素

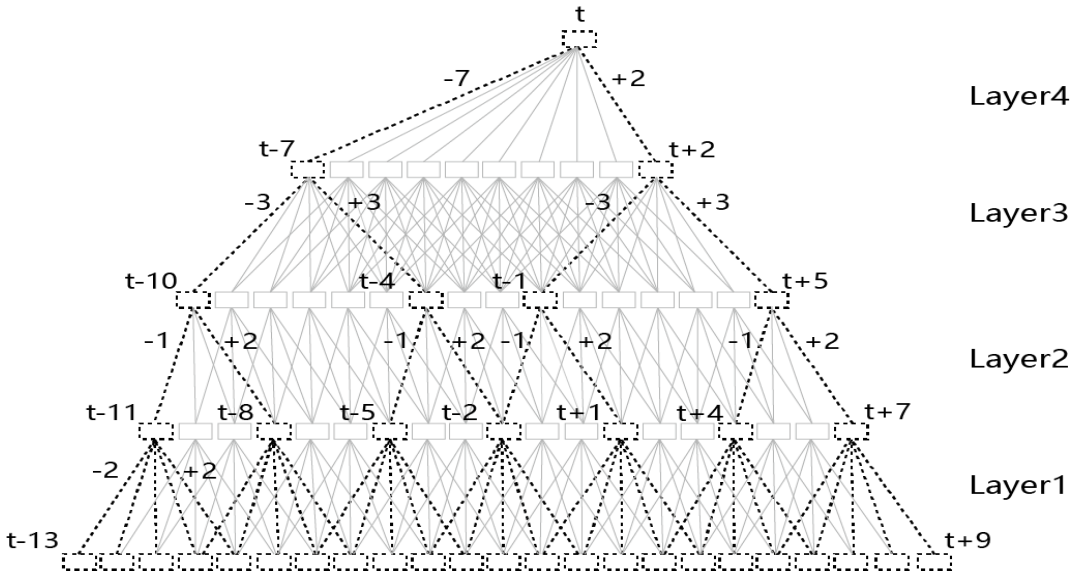
发音属性	类别数	描述
发音方式	9	发音的方式
发音位置	6	发音的位置
清浊音	2	清音/浊音
送气不送气	3	送气/不送气
舌位高低	8	舌头位置高低
舌位前后	8	舌头位置靠前/后
唇形圆展	3	嘴唇的形状

语种在发音属性特征空间中的分布



时延神经网络

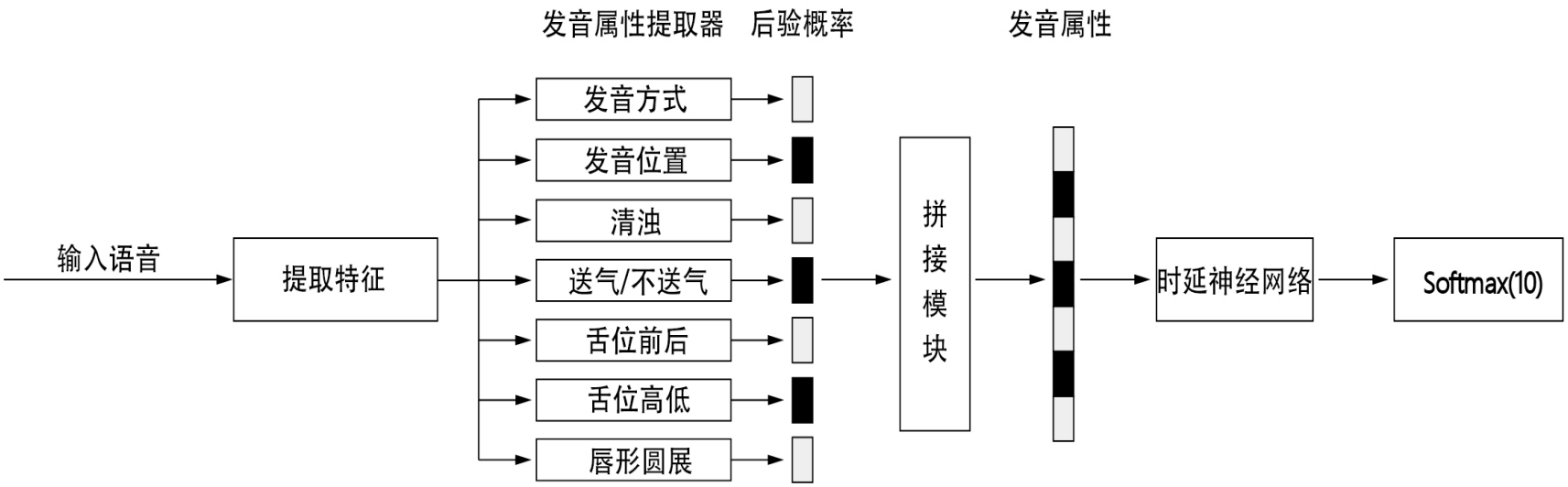
- 学习特征的上下文信息
- 建模发音属性特征空间中分布的差异



(*V. Peddinti, D. Povey et al., 2015*)

语种识别系统框图

- 基于发音属性和时延神经网络语种识别系统框图



实验设计

- 输入特征
  - MFCC
  - 深度瓶颈层特征（Deep Bottleneck, DBN）
  - 发音属性特征（AFs）
- 识别方法
  - I-vector
  - X-vector
  - TDNN

实验语料

- 发音属性&深度瓶颈层
  - AISHELL & 863
- 语种识别：AP17-OLR

AP17-OL10			AP17-训练集			AP17-测试集		
语种标识	详细描述	语音信道	话者个数	每个话者说的句子个数	句子总数	话者个数	每个话者说的句子个数	句子总数
ka-cn	哈萨克语	电话	86	50	4200	86	20	1800
ti-cn	西藏语		34	330	11100	34	50	1800
uy-id	维吾尔语		353	20	5800	353	5	1800
ct-cn	广东话		24	320	7559	6	300	1800
zh-cn	普通话		24	300	7198	6	300	1800
id-id	印尼语		24	320	7671	6	300	1800
ja-jp	日语		24	320	7662	6	300	1800
ru-ru	俄语		24	300	7190	6	300	1800
ko-kr	韩语		24	300	7196	6	300	1800
vi-vn	越南语		24	300	7200	6	300	1800

实验结果

表 2　不同方法等错误率（EER）和最低检测代价（minCavg）

特征	TDNN	ivector+cosine	xvector+cosine
MFCC	11.29(12.09)	6.22(6.87)	5.76(5.13)
DBN	7.17(6.88)	5.02(4.76)	4.43(4.53)
All-AFs	3.86(3.56)	4.56(4.32)	3.52(3.22)

系统融合结果

表 3　不同的系统融合方法等错误率（EER）和最低检测代价（minCavg）

融合系统	EER(minCavg)
MFCC−TDNN + AFs−TDNN	3.46(3.76)
DBN−TDNN +AFs−TDNN	2.95(3.21)
DBN−ivector + AFs−TDNN	2.56(2.32)
DBN−xvector + AFs−TDNN	2.21(2.36)
AFs−ivector + AFs−TDNN	2.27(2.53)
AFs−xvector + AFs−TDNN	2.14(2.01)
AFs−ivector + AFs−xvector + AFs−TDNN	1.92(1.84)

实验结果

表4　基于时延神经网络的不同发音属性特征以及不同发音属性组合特征的方法等错误率（EER）和最低检测代价（minCavg）

属性特征	EER(minCavg)
MFCC	11.29(12.09)
发音方式(MA)	9.51(10.21)
发音位置(PA)	6.53(6.14)
清浊音(VO)	10.45(10.23)
送气不送气(AS)	11.66(11.54)
舌位前后(TF)	9.82(10.28)
舌位高低(TH)	10.15(10.53)
唇形圆展(RO)	10.98(10.68)
MA + VO + AS	7.12(7.45)
MA + PA + VO + AS	5.25(4.78)
All-AFs	3.86(4.10)

总结

- 基于发音属性的时延神经网络系统性能优于基于深度瓶颈层特征的 i-vector 和 x-vector 方法。
- 系统融合可以进一步提升语种识别的性能
- **发音位置**是发音属性特征中区分性最强的特征

谢谢
请大家批评指正