多少数民族语言连续语音识别方法及应用

清华大学语音语言中心

西北民族大学民族语言信息技术教育部重点实验室

新疆大学信号与信息处理重点实验室

内容提要

- 研究背景及研究意义
- 国内外研究现状及发展动态
- 研究目标
- 研究内容
- 研究方法与步骤
- 创新点
- 研究基础与工作条件
- 计划与预算

研究背景及研究意义

中国少数民族语言概述

- 中国少数民族语言
  - 55个少数民族于,不少72种语言
  - 不少于19种文字
  - 五大语系
- 应用情况：
  - 广泛应用语言
  - 小泛围应用语言
  - 较少应用语言

http://baike.baidu.com/link?url=qOUY1ECTdAcnNwlOvtNgcRHBJA6gCdhQvOYxuuYgygC_ACENnXxXmuMyd23062c7DkaxpofRm1h28Nz0Pi-E4K

中国少数民族语言分布特征

- 高度复杂性
  - 多语系、多语族
  - 融合、衍生
- 不均衡性明显
  - 使用人口
  - 文字发展阶段
  - 资源积累
- 受主流语言影响显著
- 趋向多元化
  - 地域影响
  - 网络用语冲击

多少数民族语音识别

- 本研究关注少数民族语言的语音识别理论与方法，特别是解决多少数民族语言语音识别中的关键问题。

研究意义与价值

- 促进经济发展
- 增进民族团结
- 维护社会稳定
- 加强对外交流
- 保护语言文化

困难和挑战

- 资源稀缺
- 语言的复杂性和各异性
- 多语言融合
- 多元化

国内外研究现状

国内研究现状

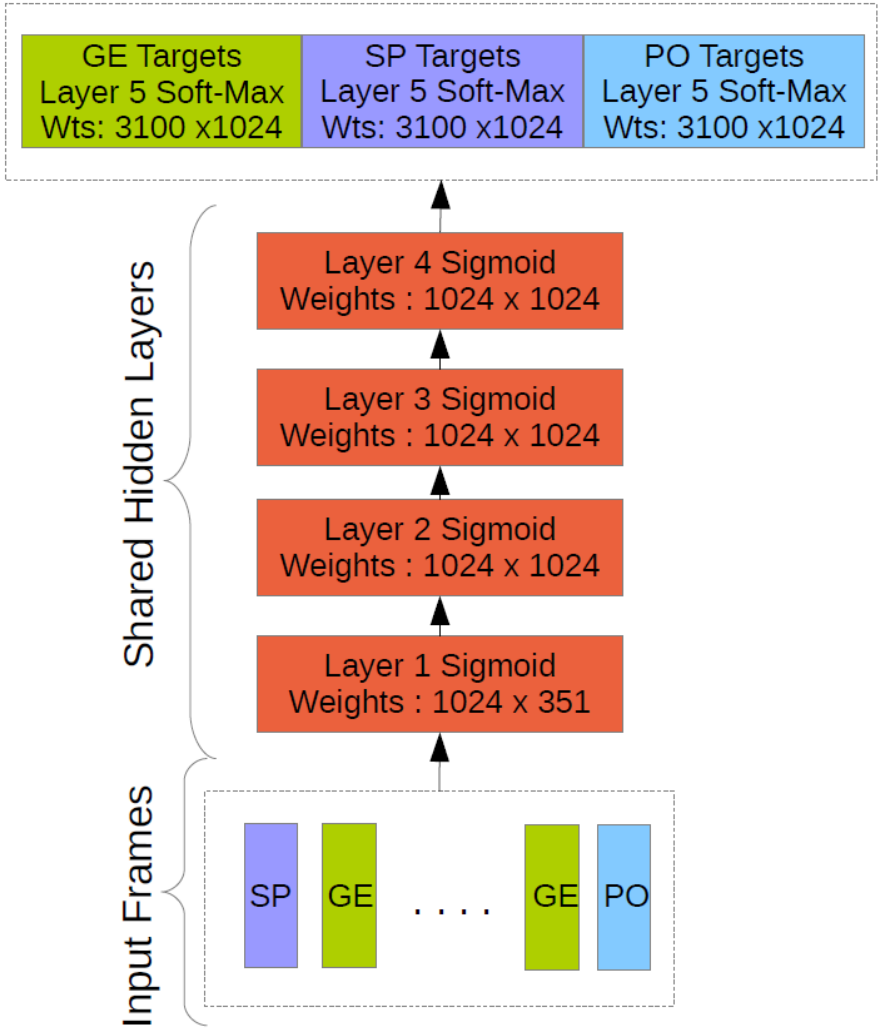
- **蒙古语语音识别**
  - 内蒙古大学:高光来,牧仁高娃等
- **维吾尔语语音识别**
  - **新疆大学：**吾守尔，艾斯卡尔，米吉提等
- **藏语语音识别**
  - **西北民族大学：**于洪志，李冠宇等
  - **中科大：**袁胜龙、郭武等

国际研究现状

- 传统多语言小语种识别
  - 音素共享
    - IPA/UPS (Schultz 2006)
    - 数据驱动(Uebler 2001, Lin 2008)
  - 特征共享
    - 后验概率特征 (Stolcke 2006)
    - Bottle Neck 特征 (Metze 2012)
  - 结构（参数）共享
    - 多语言HLDA/RDLT(Karafiát 2012)
    - MLP 低层共享(Ghoshal 2013)

国际研究现状

- 基于深度学习的多语言小语种识别
  - 多语言初始化DNN (Swietojanski 2012)
  - 共享DNN特征提取层
    - DNN-HMM(Huang 2013, Mohan2015, Sercu2016)
    - DNN Tandem(Thomas 2013, Miao2014, Chuangsuwanich 2016)
  - 结构限制
    - Stacked BN (Grezl 2014,Chuangsuwanich 2014)
    - Lowrank(Mohan2015, Sahraeian2016)

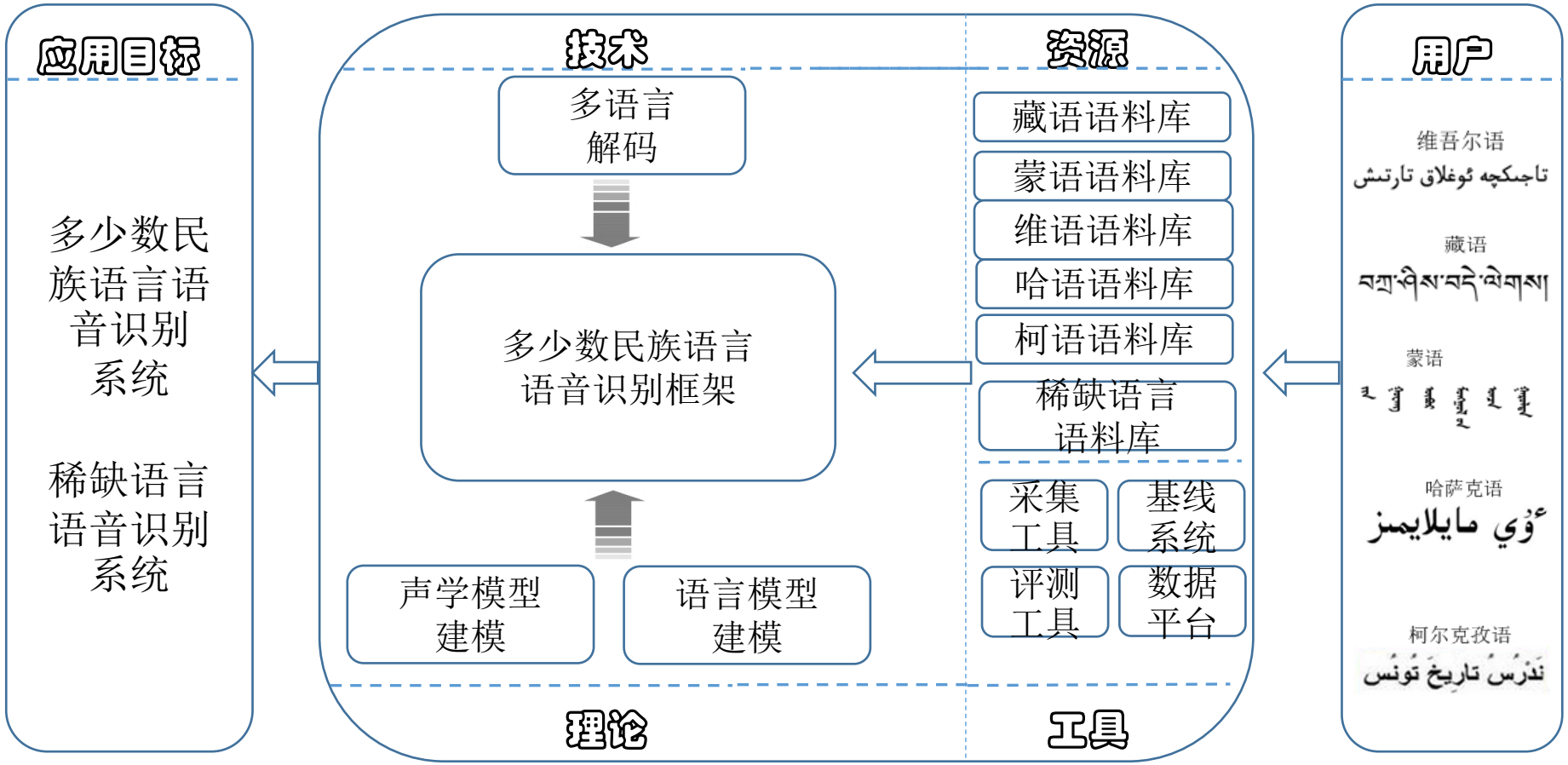


当前研究的缺陷

- 共享学习不充分
  - 集中在声学层共享，未考虑语言层
- 混合解码研究不足
  - 多语言训练研究较多，多语言识别很少
- 资源极度稀缺语言研究不足
  - 主要少数民族语言研究较多，其它少数民族语言研究较少

研究目标

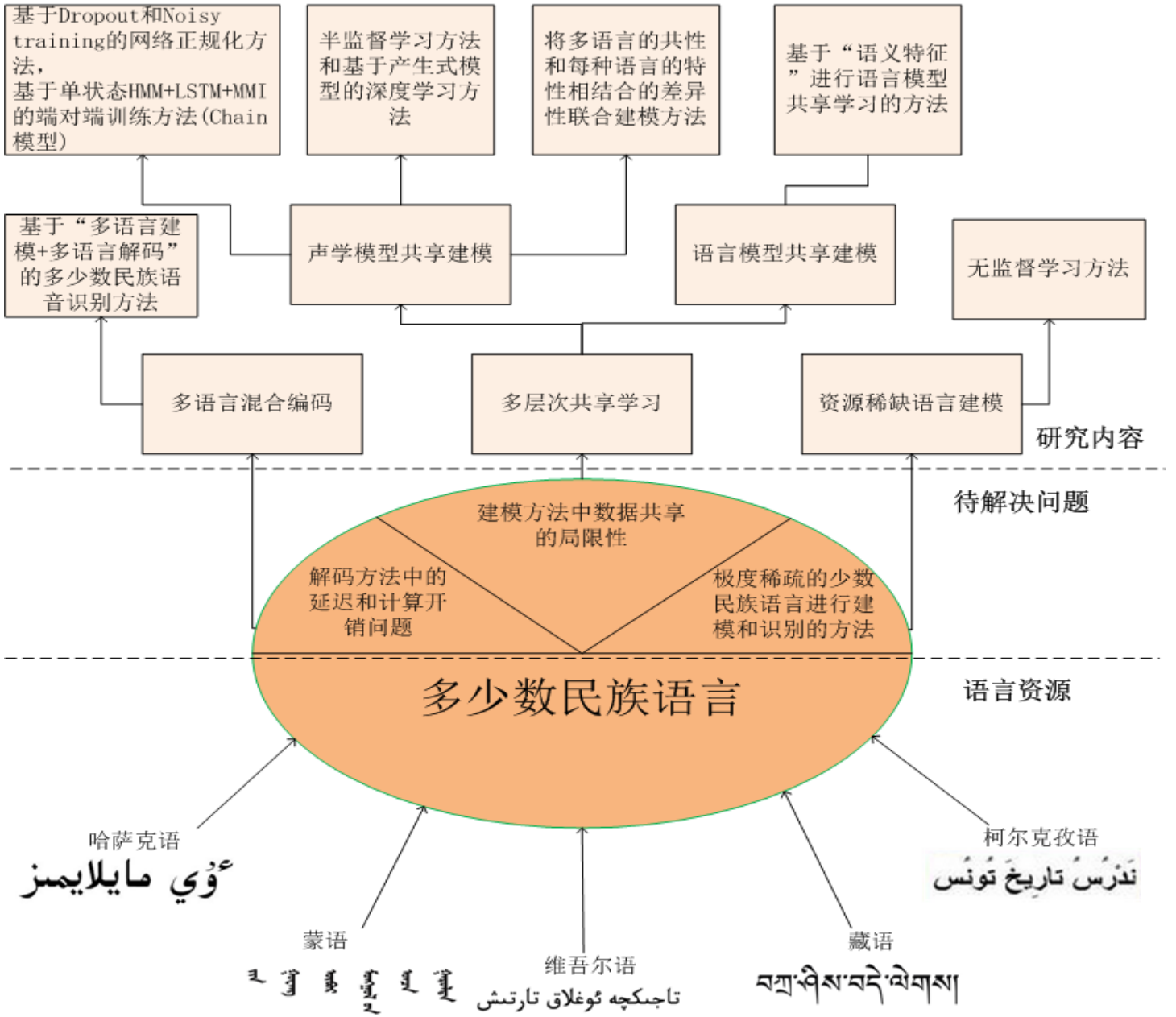
研究目标


应用目标
多少数民族语言语音识别系统
稀缺语言语音识别系统
技术
多语言解码
多少数民族语言语音识别框架
声学模型建模
语言模型建模
理论
资源
藏语语料库
蒙语语料库
维语语料库
哈语语料库
柯语语料库
稀缺语言语料库
采集工具
基线系统
评测工具
数据平台
工具
用户
维吾尔语
تاجىكچە ئۇغلاق تارتىش
藏语
བཀྲ་ཤིས་བདེ་ལེགས།
蒙语
ᠮᠣᠩᠭᠣᠯ ᠬᠡᠯᠡ
哈萨克语
ءۇي مايلايمز
柯尔克孜语
نَدْرُسُ تارِيخَ تُونُس

研究内容

研究内容

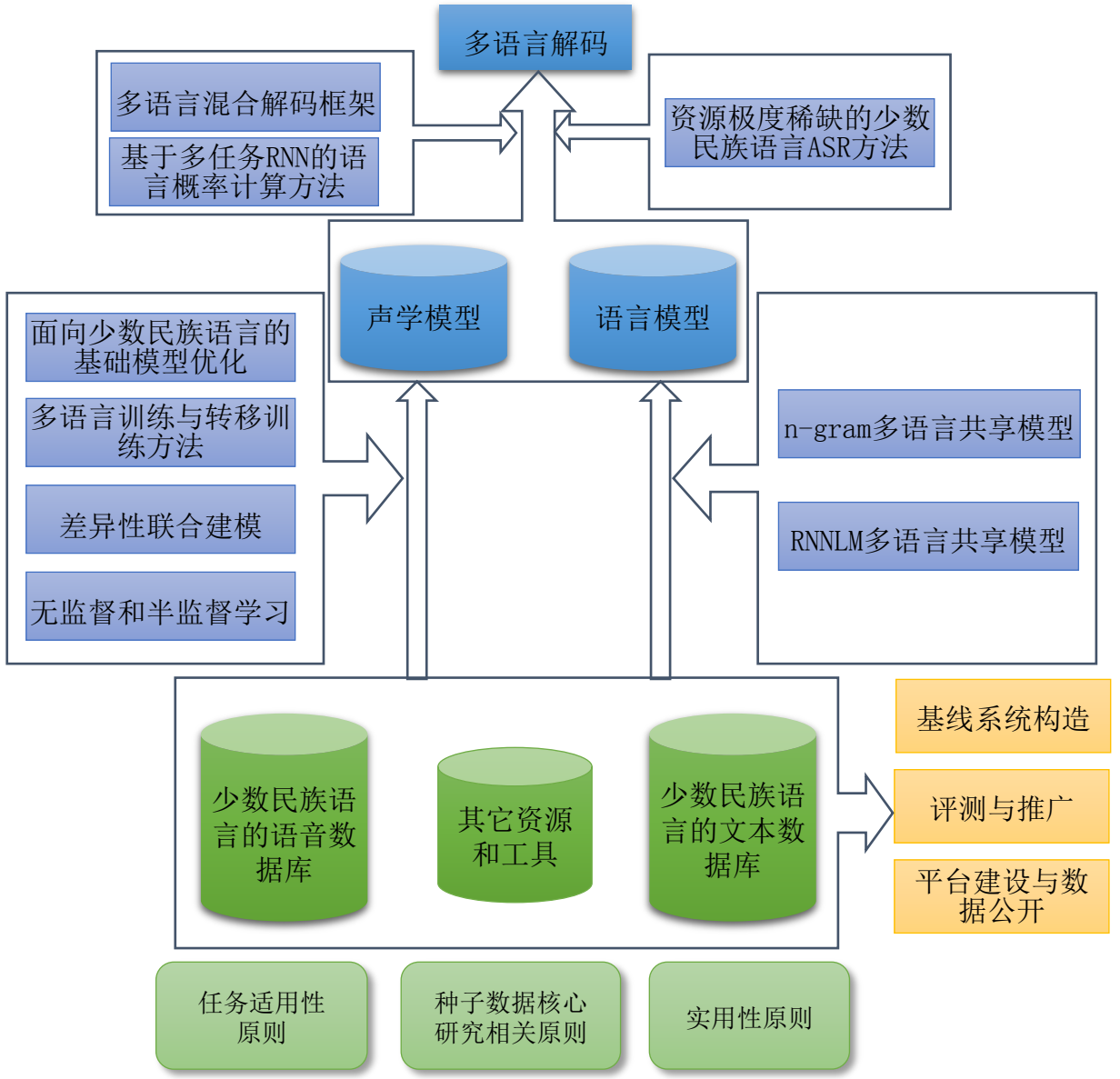
- 如何实现更有效的多语言共享建模？
  - 多层次共享学习
- 如何实现混合语言解码？
  - 增加语言约束
- 如何对资源极度稀缺语言建模？
  - 半监督学习
  - 端对端学习



研究方法与步骤

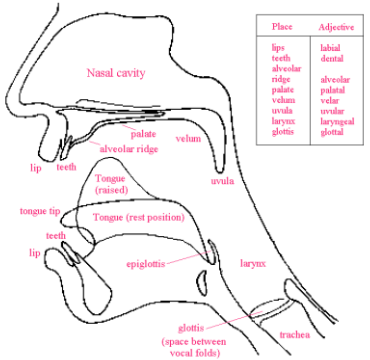
研究方法与步骤

1. 语音和语言数据库建设
2. 共享建模研究
3. 多语言混合解码方法
4. 资源稀缺语言语音识别方法



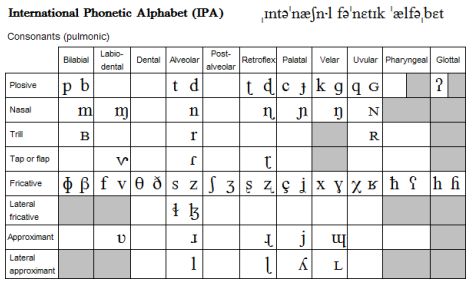
关键技术（一）多层次共享学习

- 人类语言具有多层相似性
- 当前研究仅集中于语音层



发音层

语音层

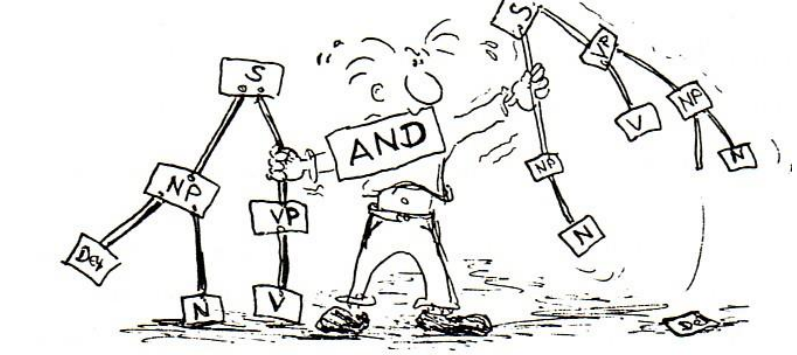


International Phonetic Alphabet (IPA) ˌɪntəˈnæʃn-l fəˈnetɪk ˈælfəˌbet

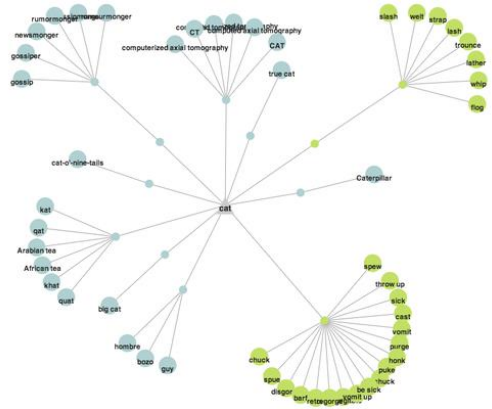
Consonants (pulmonic)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

音素层



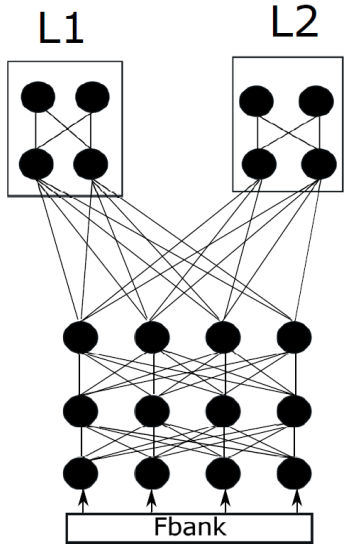
句法层



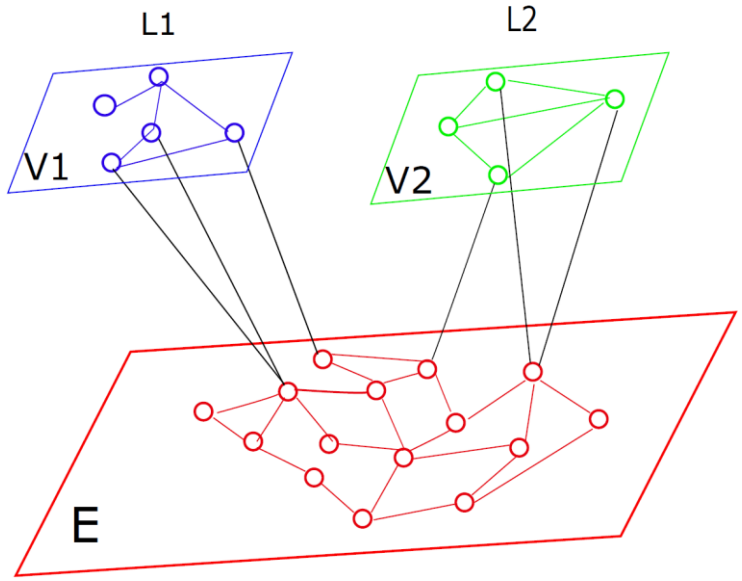
词义层

关键技术（一）多层次共享学习

- 声学模型共享：人类发音具有共性
- 语言模型共享：人类理解具有共性



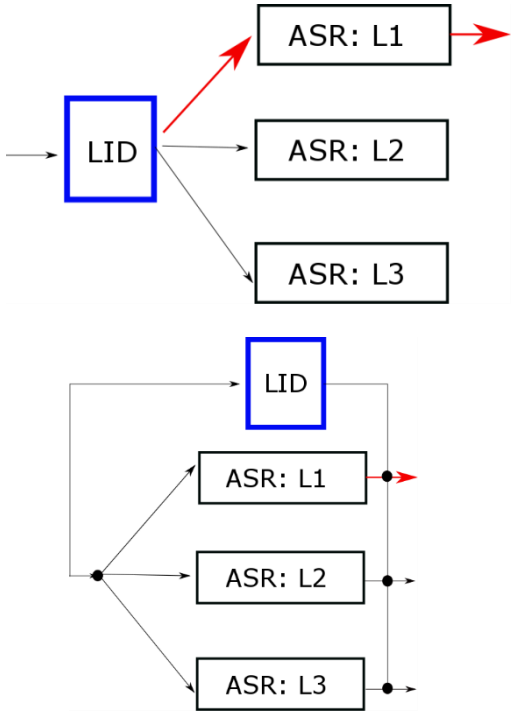
声学模型共享学习



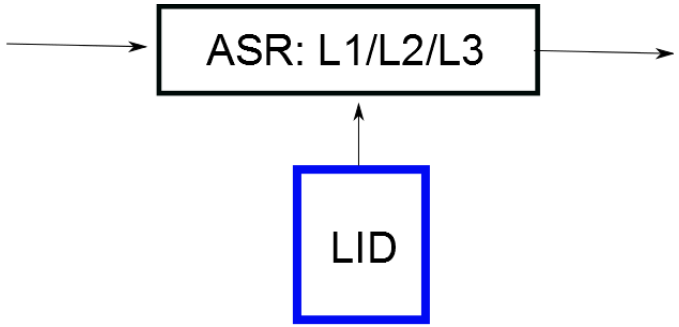
语言模型共享学习

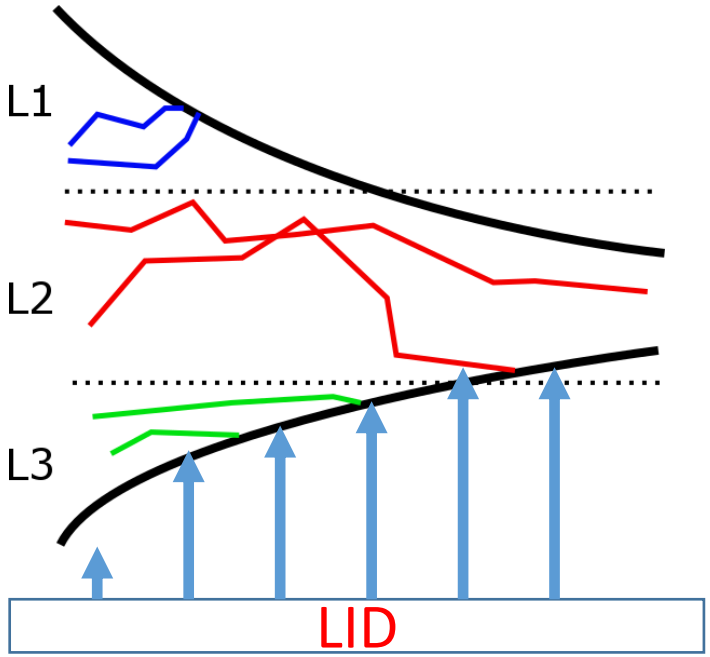
关键技术（二）多语言混合解码

- 传统方法基于语种识别
- 全空间解码任务



基于语言识别(LID)的多语言解码

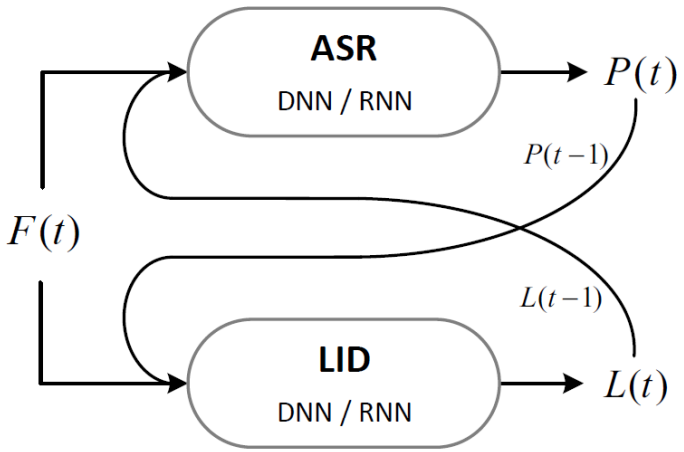




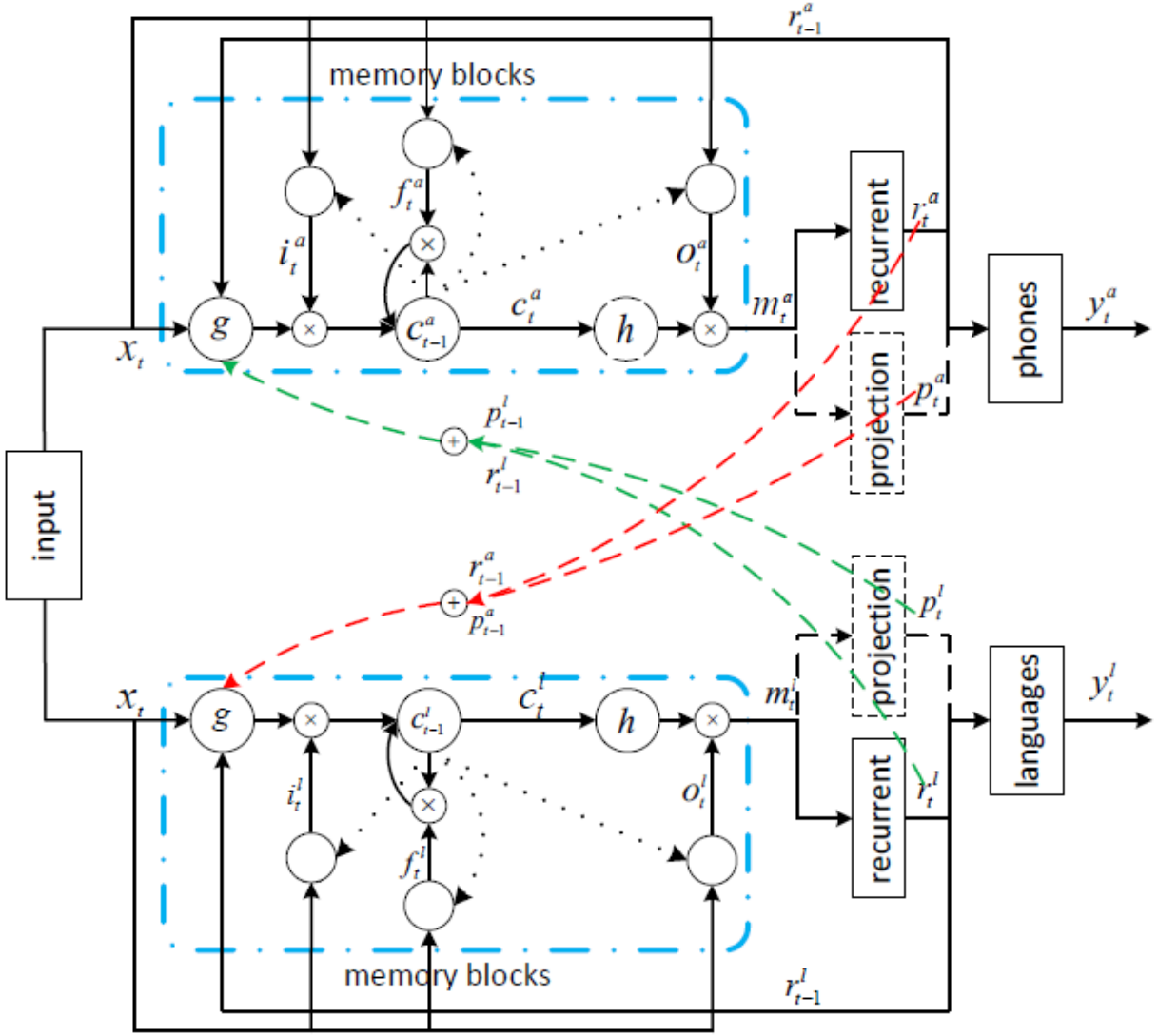
多语言混合解码

关键技术（二）多语言混合解码

- 协同解码



语音-语言协同建模

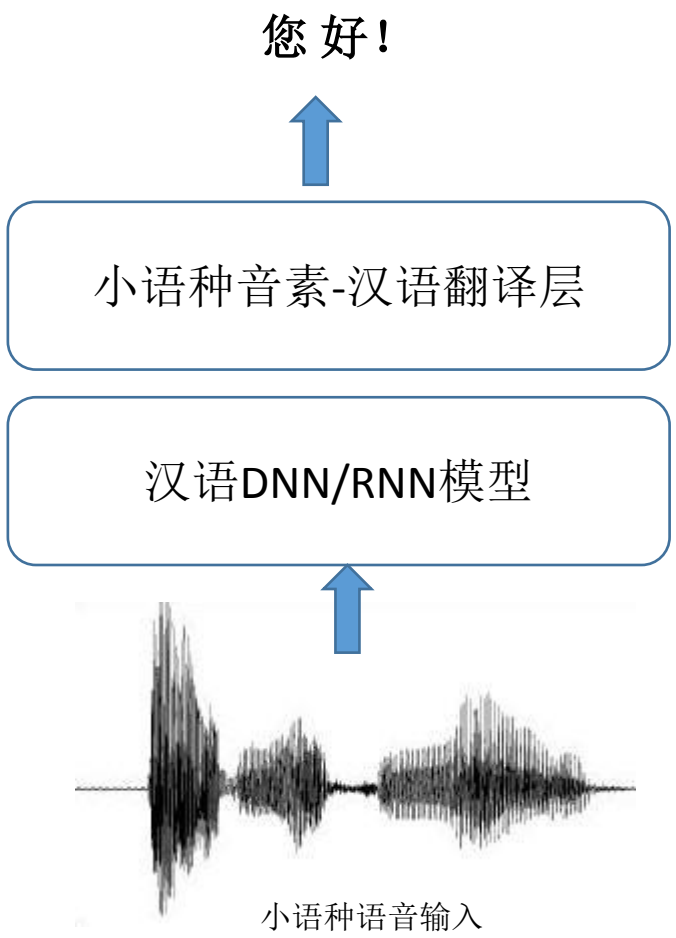


关键技术（三）资源极度稀缺语言语音识别

- 资源极度稀缺
  - 音素表
  - 词表
  - 文字
  - 标注数据
- 困难
  - 无文字，难解码
  - 无数据，难建模

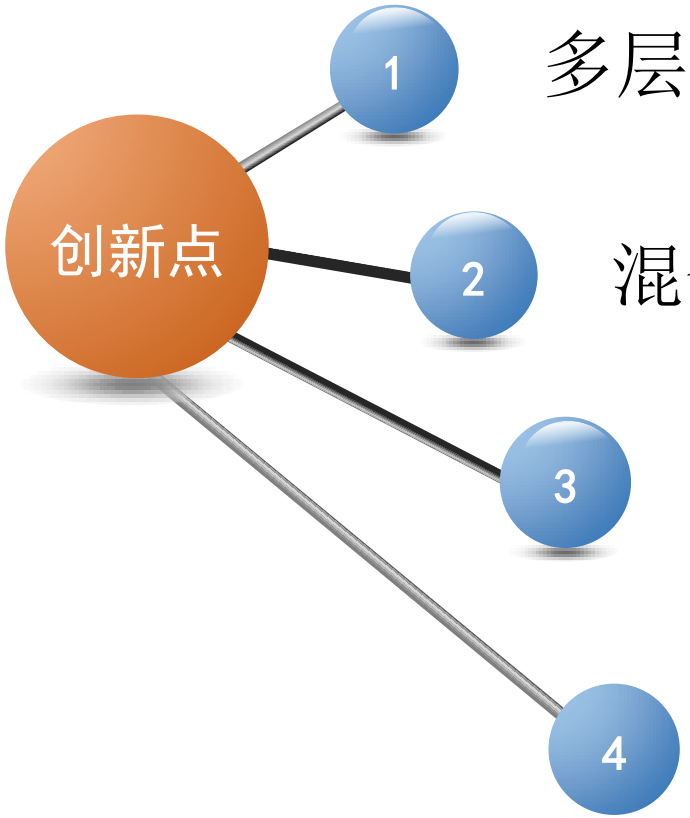
关键技术（三）资源极度稀缺语言语音识别

- 半监督建模
  - 利用相似语言
  - 实用性不强
- 端对端学习
  - 借用汉语进行标注
  - 语音识别+翻译
  - 数据稀疏问题



创新点

创新点



多层次共享训练方法

混合语言语音识别

资源极度稀缺少数民族语言语音识别

运用统一平台处理多民族语言的识别问题

工作基础

工作条件

- 工作平台
  - 三家联合申请单位具有各自实验室，拥有基础计算平台及文献资料库。
- 资源积累
  - 语音数据资源积累：汉语上万小时，维语１６０多小时，蒙语20小时，藏语150小时，哈萨克语100小时。
  - 文本数据资源积累：汉语2TB，维语5GB，藏语10GB，哈萨克语1GB，蒙语50MB。
  - 语音识别工具：THCSH30汉语和THUYG20维语语音识别系统Kaldi标准流程
- 研究成果
  - 近两年研究组在相关领域发表论文40余篇，SCI 7 篇。
  - 相关专利成果5项
- 应用成果
  - 维-汉实时语音翻译系统
  - 商用系统经验：维-汉友谊桥；灵云平台；声密宝。

研究积累

- 维语识别研究
  - Morphone 分析工具
  - 基于Morphone的Kaldi 维语识别系统
  - 基于Dark Knowledge的转移学习
- 藏语识别研究
  - 音素集、词表资源建设
  - 大词汇非特定人识别WER 17%
- 蒙语语音学研究
  - 音素集、词表建设

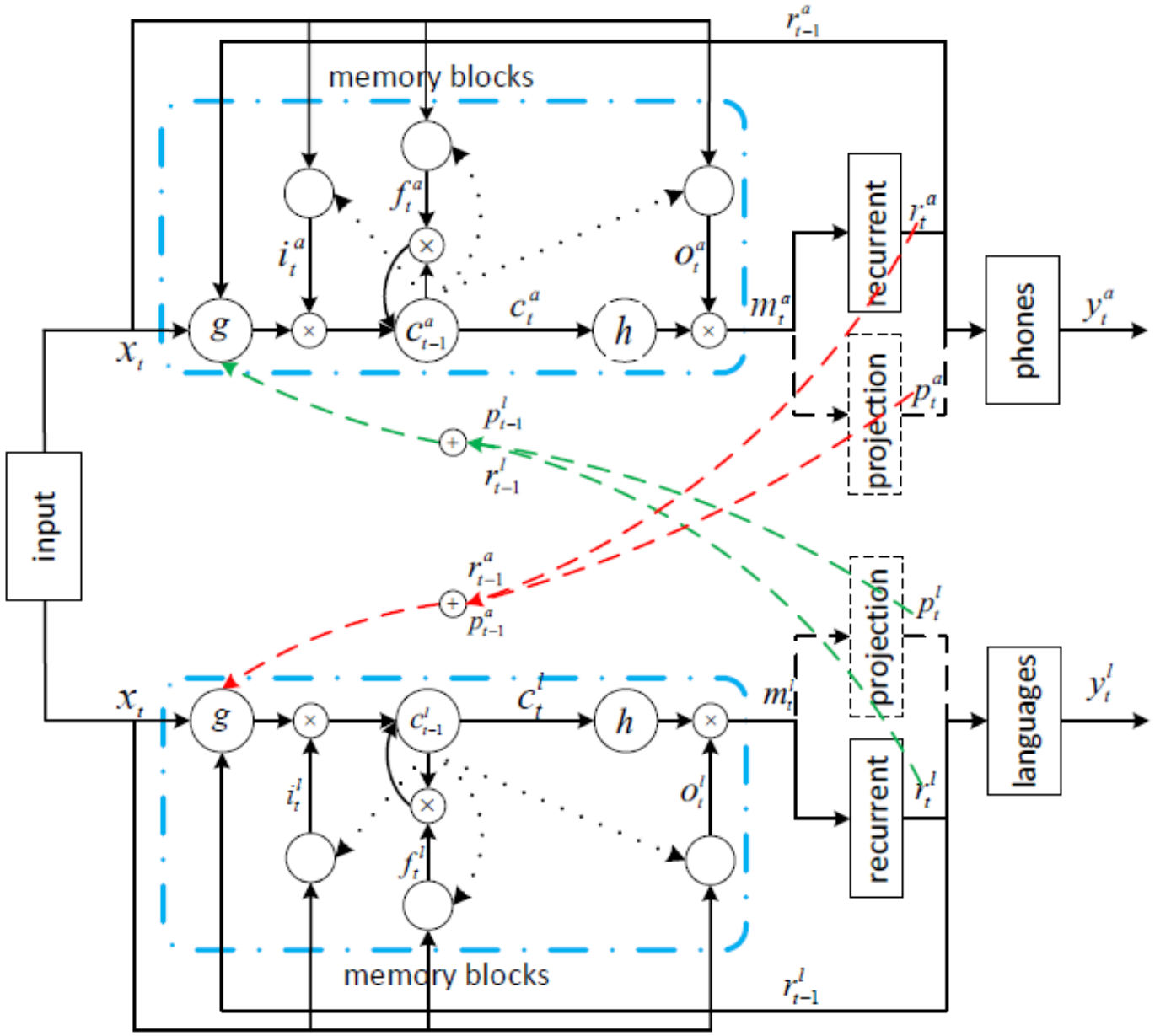
THUYG20维语识别系统

	xEnt	MPE
词系统	19.57	18.95
Morphone系统	17.40	16.58
Dark knowledge	15.20	12.25

研究积累 (二)

- 语音/语种协同建模

	英文	中文
单语言识别	12.40	23.45
混合语言识别	17.80	23.84
协同建模	17.21	23.71



进度安排

进度安排

数据库标准制定	工具建设	平台建设	数据公开与推广
数据采集	AM：基线系统优化	LM：n-gram共享	语言识别+单语言解码
	AM：半监督学习	LM：RNN共享	RNN语言-语音多任务学习
	AM：差异性联合建模		多语言混合解码
基线系统	无监督自适应	多语言标注与端对端学习	

资源建设	共享学习	混合解码	资源稀缺语言建模

任务分工

项目组织	清华大学（60%）
	西北民大（20%）
	新疆大学（20%）
资源建设	西北民大：藏语、蒙语
	新疆大学：维语、哈萨克语、柯尔克语
算法研究	清华大学 西北民大 新疆大学
平台开发	清华大学（60%）
	西北民大（20%）
	新疆大学（20%）

预期成果

预期成果（一）

- 开放的少数民族资源库
  - 维语、藏语、蒙语、哈萨克语、柯尔克孜语语音数据库
  - 相应的音素集、词典和文本数据库
  - 用于语音和语言分析的工具集
- 多少数民族语言语音研究平台
  - 开放的识别系统训练框架与流程
  - 开放的解码流程与源代码
  - 开放的资源共享平台
- 多少数民族语言语音识别平台
  - 开放的语音识别API接口
  - 开放的多语种模型库
  - 开放的多语种解码器

预期研究成果 (二）

- 在领域权威杂志发表SCI论文不少于10篇, 其它会议或杂志EI论文不少于40篇。
- 申请不少于10项发明专利。
- 培养博士生8~10名，硕士生10~15名。

经费预算

经费预算

序号	科目名称	清华大学	新疆大学	西北民族大学	总金额
1	一、项目资金	150	75	75	300
2	（一）直接费用	128	64	64	256
3	1、设备费				
4	（1）设备购置费	36			36
5	（2）设备试制费				
6	（3）设备改造与租赁费				
7	2、材料费	13	12	12	37
8	3、测试化验加工费	10			10
9	4、燃料动力费	10			10
10	5、差旅费	18	11	11	40
11	6、会议费	9	3	3	15
12	7、国际合作与交流费	10	5	5	20
13	8、出版/文献/信息传播/知识产权事务费	4	3	3	10
14	9、劳务费	13	30	30	73
15	10、专家咨询费	5			5
16	11、其他支出				
17	（二）间接费用	22	11	11	44
18	其中，绩效支出	5	3	3	11

- 欢迎批评指正！