多少数民族语言连续语音识别方法及应用

正文：参照以下提纲撰写，要求内容翔实、清晰，层次分明，标题突出。**请勿删除或改动下述提纲标题及括号中的文字。**

（一）立项依据与研究内容（5000-10000 字）：

1．项目的立项依据（研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）；

1.1 研究方向和意义

中国是个多民族国家，每个民族都有自己独特的语言。研究多民族背景下的多语言和跨语言现象，对促进语言互通互融、建设文化和谐、维护民族稳定具有重要意义。本研究关注少数民族语音识别技术，特别是多语种少数民族语言的语音识别技术，实现一个可适用于多少数民族语言的统一建模方法和识别系统，该方法和系统既可以支持资源相对丰富的少数民族语言（如维吾尔语、蒙古语），也可快速移植到资源稀缺的少数民族语言（如彝语、侗语）。同时，本项目还将建设和开放一系列数据资源、测试标准、标注平台和处理工具，推动多少数民族语言语音识别和相关领域（如信息抽取、自然语言理解等）的研究进程。

多少数民族语言语音识别的研究有助于国家均衡发展和整体振兴。与沿海经济发达地区相比，少数民族地区的经济发展相对落后，其中一个重要原因是语言障碍阻碍了与发达地区的交流。语音识别是打破语言壁垒的第一步，基于语音识别技术的少数民族语音输入法、语言翻译、多语言自动问答系统等将极大推动民族间的互通互荣。同时，语音识别技术带来一系列产业机会也将直接促进少数民族地区的经济发展，如少数民族语言呼叫中心建设、多语言口语语音评测等。

多少数民族语言语音识别的研究有助于提高社会服务质量，维护社会维定。少数民族分布广泛而分散，社会管理困难。依靠信息技术可以极大提高政府社会服务质量，其中语音识别技术作为打破语言障碍的关键一环具有重要意义。一方面，语音识别技术可以提高交流效率，改进政府服务的反应速度，另一方面，基于语言识别技术的关键信息提取技术可以对信息进行有效甄别和过滤，有利于保证社会稳定。

多少数民族语言语音识别的研究有助于提升国家影响力。当前中国正加紧布署一系列新文化交流战略，特别是新丝绸之路计划，途经众多小语种国家，打破语言壁垒是保证该计划顺利实施的重要环节。多少数民族语言语音识别系统或者可以直接应用到周边国家（如哈萨克斯坦），或者可以快速移植到相关国家的语言,促进国际间的交流合作，提升国家影响力。

1.2 我国少数民族语言概述

我国是多民族国家,包括占人口绝大多数的汉族和55个少数民族。根据2010年第六次人口普查数据，人口在百万以上的少数民族有 18 个[1]。这 55 个少数民族共使用约 72 种语言，这些语言分别属于五个语系：汉藏语系、阿尔泰语系、南岛语系、南亚语系和印欧语系，其中汉藏语系和阿尔泰语系占绝大多数。

汉藏语系[2]

包括汉语和藏缅、壮侗、苗瑶等 3 个语族。汉语在语言系属分类中相当于一个语族的地位，使用汉语的人口有 11.09 亿多；藏缅语族包括藏语、门巴语、珞巴语、僜语、嘉戎语、羌语、普米语、独龙语、土家语、彝语、傈僳语、纳西语、哈尼语、拉祜语、白语、基诺语、怒语、景颇语、阿昌语、载瓦语等，使用人口约 2200 万；壮侗语族包括壮语、布依语、傣语、侗语、水语、仫佬语、毛南语、拉珈语、黎语、仡佬语等语言，使用人口 2300 多万；苗瑶语族包括苗语、布努语、勉(瑶)语和畲语等语言，使用人口约 940 万。

汉藏语系语言一般有固定声调,很多语言的元音分长短,声母清浊音对立,一部分语言保留有复辅音声母。汉藏语系各语言以词序和虚词为表达语法意义的重要手段。词序比较固定。虚词种类较多，在句中表示各种不同的语法意义。汉藏语系语言词类上的一个特点是有量词。量词的作用主要表示事物的 单位和动作行为的量。此外，许多语言的量词还兼表事物的类别、形状、性别、级别等特征。汉藏语系语言有词的重叠形式,在许多语言里具有共同的特点,重叠的范围、作用、形式等方面。其次，各语言还有一批表示不同句式和语气的助词，一般都用于句子末尾。

阿尔泰语系[3]

---

[1] http://baike.baidu.com/view/1917.htm
[2] http://baike.baidu.com/view/18703.htm

包括蒙古、突厥、满通古斯等 3 个语族。蒙古语族包括蒙古语、达斡尔语、东乡语、东部裕固语、土族语和保安语等，使用人口约 550 万；突厥语族包括维吾尔语、哈萨克语、柯尔克孜语、乌孜别克语、塔塔尔语、撒拉语、西部裕固语和图佤语等语言，使用人口约 840 多万；满－通古斯语族包括满语、锡伯语、赫哲语、鄂温克语和鄂伦春语，使用人口约 20 万。

阿尔泰语系语言具有如下特点:

（1）具有元音和谐现象；

（2）使用后加成分为词的派生和词型变化的主要手段，粘合多个附加成分以表示多重语法意义；

（3）名词和代词有数、格等语法范畴，动词有态、时、式等语法范畴；

（4）动词在宾语之后，定语在被修饰词之前。

（5）东亚及北亚地区该语系受汉藏语系中的汉语影响明显。

阿尔泰语系各语言都是粘着语，在构词法和形态学上有很大的共同性。以在词根之后加构词附加成分为派生新词的主要手段，以在词干之后加语法粘附成分为形态变化的主要手段。构词成分和语法成分都既可以是单层的，又可以是多层的。由多个根词构成合成词的情况比较少见，在近代才逐渐增加。总之，从语音结构上、语词派生上和语法构造上看，阿尔泰语系中的各语言之间表现出高度一致性。

南岛语系

又称马来－波利尼西亚语系，高山族语言属这个语系的印度尼西亚语族，包括排湾、阿眉斯、布嫩、鲁凯、赛设特、卑南、邵、泰耶尔、赛德、邹、沙阿鲁阿、卡那卡那布等语言。

南亚语系

该语系在中国境内包括佤语、德昂语、布朗语，使用人口约 44 万。

印欧语系

在我国，该语系包括属于斯拉夫语族的俄语和属于印度－伊朗语族的塔吉克语。俄语使用人口约 1.3 万，塔吉克语使用人口约 3.3 万，共约 4.6 万人。

---

[3] http://baike.baidu.com/view/72529.htm

我国少数民族语言中，仅有 19 种有自己的文字，它们是蒙古文、藏文、维吾尔文、朝鲜文、壮文、哈萨克文、锡伯文、傣文、乌孜别克文、柯尔克孜文、塔塔尔文、俄罗斯文、彝文、纳西文、苗文、景颇文、傈僳文、拉祜文和佤文。

这些少数民族文字大都是拼音文字，在书写形式上各有不同。比如维吾尔文、哈萨克文，是从右往左书写；蒙古文、锡伯文是从上往下书写，行序从左往右。有的标点符号也不同，如藏文标点、短词和句尾用一条垂直线表示，章节段落结尾用双垂直线，全文结束用四条垂直线；蒙古文用一个点表示逗号，两点表示句号，段落末尾用四个点表示。很多少数民族在不同的历史时期使用过多种文字。如：古代维吾尔语起初用突厥字母拼写，后来用粟特文改进的回鹘字母来拼写，之后用阿拉伯字母的古代维吾尔文来拼写(哈卡尼亚文)，再后来用阿拉伯字母的察合台文，公元九—十世纪后，改用阿拉伯字母的维吾尔文来拼写。新中国成立后，曾改用西里尔字母拼写。1970 年代， 中国国务院曾为新疆的维吾尔族创制拉丁字母的维吾尔文字。现代维吾尔文的改革最近的一次是在 1983 年，使用的文字是阿拉伯字母的书写体系。

少数民族语言在使用上大致分为三种情况：

（1）蒙古、藏、维吾尔、哈萨克等民族，有大片聚居区，人口均在百万以上，文字历史也比较悠久，他们的语言除了在家庭内部、邻里亲友间使用外，还在本民族的政治、经济、文化、教育各个领域中使用，甚至在一些邻近的或者杂居在一起的其他民族中使用。

（2）彝族和傣族虽然也有成片的聚居区，有传统的文字，但是文字没有统一的规范，方言差异也较大，本民族语言在社会上的应用不如蒙古、藏、维吾尔、哈萨克等民族那么广泛。壮语和傈僳语、拉祜语、景颇语、载瓦语由于语言内部差别较大，全民族通用的共同语还没有形成，文字的使用范围也很小，语言使用情况与彝语、傣语接近。

（3）一些少数民族的语言只在本民族内部日常生活中使用，在政治生活、学校教育中往往使用其他民族的语言（主要是汉语，有些地方也使用其他少数民族语言），没有与本民族语言相一致的文字，一般使用汉字。属于这个类型的少数民族语言相当多，大约占语言总数的四分之三以上，使用人口占少数民族总人口的一半以上。

综上所述，我国少数民族语言具有如下特性：（1）高度复杂性。我国少数民族语言跨越五个语系，数十个语族，在起源、发展历史、发音特性、语法结构等

方面具有丰富的多样性，且在发展过程中融合、衍生，多语言现象极为复杂；（2）不均衡性明显。不同语言在使用人口、应用场景、文字发展阶段、语音和语言学研究、资源积累等方面存在显著差异；（3）受主流语言影响显著。我国以汉语普通话为官方语言，各少数民族语言受到汉语的显著影响。例如，汉语关键词普遍存在于各少数民族语言中，特别是在口语表达中尤其明显；（4）趋向多元化。少数民族散居化比较普遍，导致同一种语言内部标准化不足，受地域影响严重。同时，和汉语一样，少数民族语言受网络用语冲击严重，口语表达多元化趋势明显。

值得注意的是，上述几个特点既反映在声学层（如音素集，口音，协同发音等），也反应在语言层（如词表，词法，语法，语义等），这给少数民族语言的语音和语言学研究研究带来巨大挑战。

1.3 多少数民族语言语音识别的困难和挑战

少数民族语言的复杂性、不均衡性、相互融合与衍生、口语化、方言化甚至日渐网络化，给语音识别带来巨大的困难和挑战。这些挑战包括：

（1）数据资源普遍稀缺

几乎所有少数民族语言都存在资源稀缺问题（具体见 1.6）。资源稀缺性表现在语音数据、文本数据、音素集、发音词典等各个方面。即使是资源相对丰富的几个语言（如维吾尔语），数据资源的总量也很小，而且分散在各个研究机构，缺少统一的标准和规范，且很少公开。这不仅给当前项目带来很大困难，也对少数民族语音识别研究的长远发展不利。

（2）语言的复杂性和各异性形成尖锐矛盾

少数民族语言的复杂性、多样性再加上数据的稀缺性，意味着为每个少数民族语言构造独立的语音识别系统几乎是不可能的。唯一可行的是研究统一的多语言识别系统，通过共享数据资源训练共享模型，对各少数民族语言在同一框架下识别，或至少可以实现快速自适应到某种语言上。然而，少数民族语言的各异性又意味着不同语言之间的差异极为明显，难以实现共享。将这些各异性包含在一个共享系统中十分困难。

（3）多语言融合为识别系统带来挑战

不同语言的互相融合，特别是汉语对少数民族语言的影响是大势所趋，也是研究多少数民族语言语音识别系统（而非单语言识别系统）的重要动因。然而，语言融合给语音识别带来极大挑战。其一，外来语的引入带来新词，需要扩充音

素空间和词表空间，增加了解码时的混淆度，降低识别性能；其二，加入外来语需要对词表和语言模型等进行动态更新，给当前的静态解码方法带来挑战；第三，外来语在语言模型中一般不具有足够的代表性，如何将这些新词有效加入到识别系统中并不容易。从另一个角度看，外来语的加入，特别是汉语的加入，也意味着我们可以利用其它语言的一些知识来提高少数民族语言语音识别的性能。

（4）多元化带来建模上的困难

少数民族语言不论在发音和书写上都相对多元化，如地方口音的差异，口语用法与书面语用法的差异，不同教育水平的人群对汉语的接受差异，不同年龄人群对网络语、社会媒体用语等新词汇的接受差异。更有的语言基本没有标准化的发音和语法，使用语言随人群不同具有很强的随意性。这些多元化与随意性给语音识别系统带来极大挑战，特别是在数据资源稀缺的大前提下，这一挑战显得尤为严峻。

1.4 国际多语言小语种语音识别研究现状

国际上多语言小语种语音识别研究已经有近十年的历史。与国内多少数民族语音识别研究类似，研究者们关注的重点是如何解决小语种的数据稀缺问题。一个基本思路是利用人类在发音上的共性，在不同语言间共享数据。

1.4.1 传统多语言小语种语音识别

简单的多语言语音识别可以通过多个单语言识别系统平行解码结合语种识别或置信度选择实现[2][7]。本研究关注的多语音识别是指在数据稀缺前提下通过共享数据实现对数据稀缺或缺失的“小语言”进行有效建模，其基本思路是利用语言间的相似性，设计语音学上的共享结构，进而实现多语言协同训练。这一研究可分为多语言语音识别(Multilingual ASR)和跨语言语音识别(Cross-lingual ASR)两个方向，前者通过多语言数据训练通用识别系统，后者利用某一语言（通常是资源丰富语言）的数据对另一语言（通常是小语言）进行模型训练。从技术上讲，这两个研究方向没有本质区别。我们将当前多语言语音识别研究按其数据共享方式分为三种，即音素集共享、特征共享和结构（参数）共享，下面分别进行介绍。

音素共享[55]通过 IPA 或 UPS 等人工定义的多语言音素集[1][3][9]或数据驱动方法[7][26]将不同语言中的相似发音合并成等价音素类，进而用多语言语

音数据对这些等价音素类进行建模。研究表明，音素共享可以在多种模型结构上提高识别性能，如 HMM/PTM [7]，HMM/GMM [8]，半绑定隐马尔科夫模型（SCHMMs）[9][10][11],子空间混合模型（SMM）[12]和子空间高斯混合模型（SGMM）[13]。

特征共享[4][16]通过学习语言无关的特征来实现共享多语言数据。当前应用最广泛的语言无关特征多基于前向神经网络模型 (MLP)，包括后验概率（Posterior Probability）特征和 BN (Bottle Neck)特征。前者来自网络的输出向量，后者来自网络的中间层。

后验概率特征比 MFCC 等传统特征更能反应语音中的音素信息，在很多语音识别任务中得到广泛应用[4][5][14][15][18]。众多研究表明，后验概率特征，即使不经过任何变换，也可以在另一种语言中得到很好的效果[4][5]。如果经过某种变换将源语言的后验概率特征映射到目标语言，则可取得到更好的跨语言识别效果，如[1][20][21]中提出的随机音素空间映射方法。BN 特征来自 MLP 的隐藏层，与后验概率特征相比，这一特征与音素分类任务相关性较弱，因此更适合表征多语言共有的声学特性。BN 特征在多语言语音识别任务中取得了很好的效果[16][17]。不论是后验概率特征还是 BN 特征，语音识别系统都是基于 Tandem 结构[2][19][23]，即将原始语音特征经过 MLP 提取高级特征(后验概率或 BN 特征)，再基于该特征进行声学建模，如隐马尔可夫模型/混合高斯模型（HMM/GMM）。

结构或参数共享是指通过共享模型结构和参数实现多语言间的数据。例如[6]中提出利用多语言数据训练特征变换矩阵（HLDA 和 RDLT）。当前应用最广泛的是基于 MLP/HMM Hybrid 结构的共享方法。 在该方法中，HMM 模型中每个状态似然函数值由 MLP 输出的后验概率计算得到。基于 MLP 的分层结构，可以通过共享 MLP 底层结点的方法实现识别系统的参数共享[22][24][25]。

1.4.2 基于深度学习的多语言小语种语音识别

近年来，深度神经网络（DNN）在语音识别领域取得了巨大进展。DNN 是拥有多个隐藏层的 MLP，在语音信号学习方面具有很多优势，包括：（1）多层结构有利于学习语音信号中的层次性特征； （2）DNN 网络训练的目标是音素音区分性最大，这一优化目标和语音识别任务直接相关；（3）高度紧凑共享的结构有利于多条件、多任务协同学习。其中第（3）点使得 DNN 在多语言建模方面具有天然优势。

作为浅层 MLP 模型的扩展，DNN 同样可以用于上述特征共享（Tandem 结构）和结构共享（Hybrid 结构）两种多语言建模方案中。与基于浅层 MLP 的方法相比，基于 DNN 的方法有更高的灵活度。因为声音信号中既包含人类语言的共性也包含每种语言的特性，DNN 的层次学习可以帮助我们将共性和特性区分开，在共性部分（底层）共享，在特性部分（高层）独立建模。

早期基于 DNN 的跨语言研究利用多语言数据初始化单语言的 DNN 网络，使得目标语言的 DNN 模型训练更容易[56]。 随后的研究将该方法应用到 DNN-HMM 混合模型和 Tandem 模型中。这在两种方法中，不同语言共享 DNN 的隐藏层，同时每种语言拥有各自的输出层[57][58][59]。DNN-HMM 方法用该模型计算 HMM 的状态似然函数， Tandem 方法用该模型提取语言无关特征[60][61][62]。

有些研究者进一步提出将 DNN-HMM 模型和 Tandem 结构结合，先从多语言 DNN 模型得到 BN 特征，再将 BN 特征和原始特征(MFCC)拼接起来训练 DNN-HMM 模型[63]。与之类似的一种方法是对每种语言分别提取 BN 特征，将这些多语言的 BN 特征拼接起来，训练 DNN-HMM 模型[64]。

1.5 我国多少数民族语言语音识别研究现状

和国际同行相比，我国少数民族语言语音识别研究差距还比较大，不同少数民族语言的研究基本上处于独立状态，多少数民族语言的研究还不多。我们给出几种主要少数民族语言语音识别的研究现状。这些语言使用人群相对集中，特点比较显著，且代表我国两个主要语系（汉藏语系和阿尔泰语系），因此在本项目中作为典型少数民族语言进行研究。

蒙古语语音识别：内蒙古大学计算机学院的高光来教授领导的团队曾先后做过蒙古语语音识别的相关研究[27][28][29][30]。他们的论文在对蒙古语语音声学模型的设计与优化、蒙古语语言模型的建立和优化等方面做了研究。例如单音素和三音素模型的比较，自底向上策略和自顶向下策略的比较，CHMM 模型的混合分量的增加对性能的影响，CHMM 模型和多数据流 SCHMM 模型的比较，二元文法和三元文法语言模型的比较，各种语言模型的平滑算法的比较等。内蒙古大学蒙古学学院的牧仁高娃博士对蒙古语语音识别方面也做了很多研究[31]，例如建立了蒙古语语音数据库和相应的标注文本库（该库包含 3299 条句子，共 30 多个小时的语音以及相应的标注读音文本），研制了国内首个蒙古语自动发音标注工具，建立了基础的蒙古语语音识别系统。以上的研究工作都取得了一定的成就，

在蒙古语语音识别方面做出了有益的尝试。

维吾尔语语音识别：维吾尔语是我国使用人数较多的少数民族语言，文献[32][33][34]基于 HTK 初步研究了维吾尔语连续语音识别。文献[35] 基于 DTW 方法实现了小词汇量特定人维吾尔语孤立词语音识别系统。文献[36][37]结合维吾尔语的语言特征，完成了维吾尔语基础文本设计、语音语料库录制、以音素作为基元的声学模型训练、基于字母的语言模型设计等工作。文献中统计了维吾尔语 32 个音素的识别率，给出了容易混淆的音素及其根源分析，为进一步提高识别率奠定了基础。文献[38][39][40]建立了基于 HMM 的大词汇量连续语音识别系统（LVCSR），使用了 150 小时标准语音库和 60 多万句子的文本库。从维吾尔语声学及形态学角度分析了构词特征，建立了词、词素、音节等多个层次为基础的识别系统。为了降低单词错误率（WER），提出了区分性训练方法，从词-词素两层语言粒度中重新组合一个中间层语言粒度，充分利用了语言的特点。文中设计了基于一系列特征和评估函数进行粒度单元的优化选择。文献[41]用拼接词素的方法选择词条，取得了显著的性能提高。文献[42][43]提出了适合维吾尔语的基于音节的语言模型建模方法，引入了最大匹配分词算法评价音节语言模型在大词汇连续语音识别任务中的性能。文献[44]在训练维吾尔语声学模型利用了 Tandem 特征，并基于最小音素错误区分性训练准则训练声学模型。以上工作表明，维吾尔语语音识别技术已经初具规模，但系统性能还需要提高。

藏语语音识别：藏语语音识别起步较晚，目前还处于初期研究阶段。2005 年，武光利、于洪志、戴玉刚首先对藏语语音合成的语音信号进行频谱转换与分析[45]。同年，李洪波、于洪志研究了基于藏语语音学知识的语音端点检测[46]，实现了安多藏语语音识别的端点检测方法，形成具有藏语语音特色的算法。2006 年李洪波、于洪志对藏语语音识别的预处理技术进行研究[47]， 提出将音素作为藏文语音识别的语音基元，并在噪音背景下采用倒谱特征进行语音端点检测。2007 年，于洪志、李永宏、索南楞次、仁青多杰、李毛吉研究探讨了安多藏语单音节声学参数数据库的建设[48]，建立了单音节文本库、语音库和声学参数数据库等，并对安多藏语的单音节声学特征进行了系统的研究。2007 年，刘静萍、德熙嘉措研究了安多藏语辅音识别技术[49]，研究了针对藏区安多方言的小词汇量孤立词的识别，他们采用 LPCC 系数作为特征参数，采用动态时间规整的模式匹配技术提高识别率。2007 年，武光利、于洪志对基于短时能频值的藏语语音的端点检测技术进行了研究[50]，同时，李洪波和于洪志提出了基于 LFCC 和 MFCC

的藏语语音端点检测算法[51]。根据藏语元音/辅音频谱特性差异，对语音信号分高/低频带后再分别处理。2009 年，李勇、于洪志、达哇彭措研究了藏语语音韵律参数提取[52]，他们使用关联规则的数据挖掘方法对韵律特征中的基频参数进行知识发现，获得有用的藏语韵律规则。2009 年，姚徐、李永宏、单广荣、于洪志研究了藏语孤立词语音识别系统[53]。他们采用 MFCC 作为特征参数，形成语音模板库，采用 DTW 技术实现语音识别系统。2012 年，李冠宇、孟猛研究了基于 HTK 的藏语拉萨话大语表连续语音识别方法，建立了基于上下文的 GMM-HMM 声学模型，实现了特定人和非特定人的藏语拉萨话大词表连续语音识别[66][67]。2015 年袁胜龙、郭武、戴礼荣等对藏语大词汇电话连续语音识别进行了研究，在基于深度神经网络(DNN)的声学模型建模中，针对数据稀疏的问题，提出采用大语种数据训练的 DNN 对目标语言的 DNN 模型进行初始化的策略[68]。

虽然藏语语音识别技术的起步比较晚，但是随着民族语音信息处理技术的不断发展，藏语语音识别技术已经在预处理、端点检测、语音韵律参数提取、语音数据库的建立及藏语孤立词语音识别系统、藏语连续语音识别系统等方面取得了一系列成果。然而，在大规模语料库的建立、关键词提取及语音信息检索等技术上，还缺少足够的研究，大词表连续语音识别准确率较低。

其它少数民族语音识别：文献[69]探讨了 DTW 和 EDTW 在彝语孤立词识别中的应用，并提出了两种新算法:基于音节个数的动态时间规整算法（SDTW）和基于音节个数的高效动态时间规整算法(SEDTW)，对特定的彝语语音信号进行识别。文献[70]以音素为基本识别单元，利用 HTK 工具提取 39 维 MFCC 语音特征参数，构建 HMM 模型，采用 Viterbi 算法进行模型训练和匹配来实现白族语音的识别，识别准确率达到 93.3%。文献[71]以纳西语口音为实例，利用发音变异规律实现了由标准普通话识别器到民族语口音普通话识别器的变换。

1.6 当前研究的局限性

由上述国际和国内研究现状可见，基于 DNN 的多语言特征共享方法是当前多语言小语种研究的主流方法。该方法基于不同语言在发音上的共性，利用深度学习方法对特征的抽象能力，提取语言无关特征，建立更鲁棒的声学模型。这一特征共享方法较之前的音素共享方法更自然，也更符合数据的实际分布特性，因而得到广泛认可，成为当前多语言小语种研究的标准方法，在少数民族语音识别上也有应用，如[68]。

然而，不论国内还是国外，当前对多语言小语种的研究还相当有限，特别是能考虑到我国少数民族语言特性的多少数民族语音识别研究还不多。总结起来，当前研究的局限性主要包括如下三点：

(1) 建模上数据共享不充分：当前研究主要集中在声学层，基本方法是共享不同语言的声学特征。我们已经提到，多少数民族语音识别的困难不仅在于声学上的数据稀缺，而在于声学、词法、语言等多层次的复杂性、各异性、多元化。这些仅依靠现有的声学特征共享是无法解决的。因此，我们认为当前集中在声学层的研究方法缺乏足够的层次覆盖性，未能充分利用有限的数据资源。

(2) 识别上混合解码研究不足：当前多少数民族语音识别研究大多数是“多语言训练”，而非“多语言识别”。这一方法在模型训练（主要是声学模型）中利用多语言资源，但在实际解码时依然是单语言解码。为了得到多语言识别的效果，往往需要一个语种识别做为前端并结合并行解码。这种方法事实上是用延迟的增加和计算开销的加大规避了多语言识别时搜索空间膨胀和路径混淆带来的困难，而后者是处理多语言融合现象必须要面对的问题。不论从研究角度还是实用角度，我们都需要一个基于多语言混合解码的多少数民族语言语音识别方法。

(3) 对资源极度稀缺语言的研究不足：当前研究多集中在几个主要语言上（如维吾尔语、蒙语、藏语等）。这些语言固然也存在数据资源不足的问题，但其语音学和语言学研究已经展开，语言资源和工具相对充足，识别系统构造没有实质困难，性能提高只是数据积累问题。特别是，这几个语言覆盖的人口众多且较为集中，网络资讯也比较发达，语音数据和文本数据的采集都比较容易。真正的困难在于我国少数民族语言数量庞大，绝大多数语言处于资源极度稀缺状态，使用人口分散，数据采集困难，语言学研究不足，甚至连文字都没有。如何利用已有资源为这些“主要少数民族语言”外的“极少数民族语言”建设语音识别系统，当前研究极少涉及。

参考文献

[1]David IMSENG ,Multilingual speech recognition A posterior based approach,2013

[2]D. Imseng, H. Bourlard, andM.Magimai.-Doss. Towards mixed language speech recognition systems. In Proceedings of Interspeech, 278–281, 2010

[3]Lin, Hui, et al. "A study on multilingual acoustic modeling for large vocabulary ASR." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.

[4]A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), 321–324,2006

[5]L. Tòth, J. Frankel, G. Gosztolya, and S. King. Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian. In Proceedings of Interspeech, , 2695–2698,2008

[6]Karafiát M, Janda M, Cernocky J, et al. Region dependent linear transforms in multilingual speech recognition，Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 4885-4888,2012
[7]Uebler U. Multilingual speech recognition in seven languages. Speech communication, 35(1): 53-69,2001
[8]N.T. Vu, F. Kraus, and T. Schultz. Rapid building of an asr system for under-resourced languages based on multilingual unsupervised training. In Proceedings of Interspeech, 3145–3148,2011.
[9]T. Schultz and A.Waibel. Language independent and language adaptive acoustic modeling for speech recognition. Speech Communication, 35:31–51,2001
[10]J. Köhler. Multilingual phone models for vocabulary-independent speech recognition tasks. Speech Communication, 35:21–30,2001
[11]T. Niesler. Language-dependent state clustering for multilingual acoustic modelling. Speech Communication, 49:453–463, 2007
[12]Yajie Miao, Florian Metze, Alex Waibel .Subspace mixture model for low-resource speech recognition in cross-lingual settings, In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),4334–4337, 26-31,2013.
[13]L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4334–4337, 2010.
[14]S. Dupont, C. Ris, O. Deroo, and S. Poitoux. Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In Proceedings of the Automatic Speech Recognition and UnderstandingWorkshop (ASRU), 29–34, 2005.
[15]S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F.Mana. On the use of a multilingual neural network front-end. In Proceedings of Interspeech, 2711–2714,2008.
[16]Vu, N. T., Metze, F., & Schultz, T. (2012, May). Multilingual bottle-neck features and its application for under-resourced languages. In SLTU (pp. 90-93).
[17]Vu, N. T., Breiter, W., Metze, F., & Schultz, T. (2012). An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance. In Proc. Interspeech (pp. 2586-2589).
[18]Thomas, S. , Baltimore, MD,Ganapathy, S. ; Hermansky, H. MULTILINGUAL MLP FEATURES FOR LOW-RESOURCE LVCSR SYSTEMS,ICASSP ,4269-4272,2012
[19]H. Hermansky, D. P. W. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMMsystems. In Proceedings of IEEE International Conference on Acoustics,Speech, and Signal Processing (ICASSP), 1635–1638, 2000
[20]D. Imseng, H. Bourlard, and P. N. Garner. Using KL-divergence and multilingual information to improve ASR for under-resourced languages. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4869–4872, 2012b
[21]Imseng D, Bourlard H. Speaker adaptive kullback-leibler divergence based hidden markov models, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7913-7917.
[22]Ghoshal A, Swietojanski P, Renals S. Multilingual training of deep neural networks, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7319-7323.
[23]Sequence classification using high-level features extracted from deep neural networks, ICASSP ,7319-7323,2014
[24]Dongpeng Chen, Mak, B.,Cheung-Chi Leung,Sivadas, S. Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition,ICASSP,5592-5596,2014
[25]Heigold, G. , Vanhoucke, V. ; Senior, A. ,Nguyen, P. MULTILINGUAL ACOUSTIC MODELS USING DISTRIBUTED DEEP NEURAL NETWORKS ,Acoustics,Speech and Signal Processing (ICASSP),8619-8623, 2013
[26]Hui Lin, Li Deng, Jasha Droppo, Dong Yu, and Alex Acero . Learning Methods in Multilingual Speech Recognition,NIPS Workshop, Whistler, BC,Canada ,2008
[27]包世恩. 蒙古语非特定人大词汇量连续语音识别系统的研究与实现，内蒙古大学计算机学院硕士毕业论文, 2005 年.
[28]哈斯其劳. 面向语音识别的蒙古语语音模型的研究，内蒙古大学计算机学院硕士毕业论文, 2008 年.
[29]艾霞. 面向语音识别的蒙古语语言模型的研究，内蒙古大学计算机学院硕士毕业论文, 2007 年.
[30]飞龙. 蒙古语语音识别系统的研究与优化，内蒙古大学计算机学院硕士毕业论文, 2009 年.
[31]牧仁高娃. 蒙古语语音识别相关问题研究，内蒙古大学计算机学院博士毕业论文, 2013 年.
[32]蔡琴，吾守尔·斯拉木，基于 HTK 的维吾尔语连续数字语音识别,现代计算机,2007(4)
[33]那斯尔江·吐尔逊，吾守尔·斯拉木，陶梅，基于 HTK 的维吾尔语连续语音识别研究，第七届中文信息处理国际会议,武汉,2007
[34]冯丽娟，吾守尔·斯拉木，维吾尔语连续语音识别技术研究，现代计算机,2010(1)
[35]诺明花，吾守尔，维吾尔语孤立词和连续数字语音识别设计与实现，第十一届全国民族语言文字信息学术研讨会，西双版纳,2007
[36]米日古力·阿布都热素，米吉提·阿不力米提等，基于 HTK 的维吾尔语连续音素识别技术研究，计算机工程与应用,2013(22)

[37]米日古力·阿布都热素，艾克白尔·帕塔尔，艾斯卡尔·艾木都拉，基于电话语料的维吾尔连续音素识别，通信技术,2012(7)
[38]Mijit Ablimit, Tatsuya Kawahara, Askar Hamdulla, Lexicon Optimization for Automatic Speech Recognitionbased on Discriminative Learning, APSIPA ASC, Xi'an, 2011
[39]Mijit Ablimit, Tatsuya Kawahara, Askar Hamdulla, DISCRIMINATIVE APPROACH TO LEXICAL ENTRY SELECTION FOR AUTOMATIC SPEECH RECOGNITION OF AGGLUTINATIVE LANGUAGE, ICASSP, Kyoto, 2012
[40]Mijit Ablimit, Tatsuya Kawahar, Askar Hamdulla, Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language, Speech Communication, Volume 60, 2014
[41]Mijit Ablimit, Askar Hamdulla , Tatsuya Kawahara, Morpheme Concatenation Approach in Language Modeling for Large-Vocabulary Uyghur Speech Recognition, Oriental COCOSDA, Hsinchu, 2011
[42]努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木，热依曼·吐尔逊，基于音节的维吾尔语大词汇连续语音识别系统，清华大学学报,2013(6)
[43]努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木，热依曼·吐尔逊，基于音节的维吾尔语大词汇连续语音识别系统，第十二届全国人机语音通讯学术会议，贵阳, 2013
[44]努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木，维吾尔语连续语音识别声学模型优化研究，计算机工程与应用, 2013(2)
[45]武光利,于洪志,戴玉刚.藏语语音合成系统中语音信号的频谱转换与分析[J]. 西北民族大学学报（自然科学版）,2005，26（59）.
[46]李洪波,于洪志.基于藏语语音学知识的语音端点检测研究.第七届中文信息处理国际会议论文集,2007.
[47]李洪波,于洪志.藏语语音识别的预处理研究.中国中文信息学会二十五周年学术会议论文集,2006.
[48]于洪志,李永宏,索南楞次,仁青多杰,李毛吉.安多藏语单音节声学参数数据库研究探讨.第十一届全国民族语言文字信息学术研讨会论文集,2007.
[49]刘静萍,德熙嘉措.安多藏语辅音识别技术.民族语言文字信息技术研究,第十一届全国民族语言文字信息学术研讨会论文集,2007.
[50]武光利,于洪志.基于短时能频值的藏语语音的端点检测的研究．促进中西部发展,声学学术交流会论文集,2007.
[51]李洪波,于洪志.基于 LFCC 和 MFCC 的藏语语音端点检测算法,2007.
[52]李勇,于洪志,达哇彭措.基于关联规则的藏语语音韵律参数提取．微计算机信息（管控一体化），2009.25（2- 3）：255- 257.
[53]姚徐,李永宏,单广荣,于洪志.藏语孤立词语音识别系统.西北民族大学学报（自然科学报）,2009，30（73）
[54]D. Yu, K. Yao, H. Su, G. Li, and F. Seide. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP). 2013.
[55]T. Schultz. Multilingual acoustic modeling. In Multilingual Speech Processing, pages 71–122. Academic Press, 2006.
[56]Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Unsupervised cross-lingual knowledge transfer DNN-based LVCSR," in Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012, pp. 246–251.
[57]Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7304–7308.
[58]Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 8619–8623.
[59]Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7319–7323.
[60]Karel Vesely, Martin Karafiat, Frantisek Grezl, Marcel Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012, pp. 336–341.
[61]Stephan Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 6704–6708.
[62]Zoltan Tuske, Joel Pinto, Daniel Willett, and Ralf Schluter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7349–7353.
[63]Peter Bell, Pawel Swietojanski, and Steve Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 6975–6979.
[64]Jonas Gehring, Quoc Bao Nguyen, Florian Metze, and Alex Waibel, "DNN acoustic modeling with modular multi-lingual feature extraction networks," in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013, pp. 344–349.
[65]Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, Max Welling, Semi-supervised Learning with Deep

Generative Models, NIPS2015.
[66]李冠宇,于洪志，李永宏，马宁. 基于决策树的藏语拉萨话三音子模型[J]，计算机工程与科学. 2013,9. 第35 卷第 9 期：146-150
[67]李冠宇，孟猛. 藏语拉萨话大词表连续语音识别声学模型研究[J],计算机工程,2012,3(5)，189-191
[68]袁胜龙，郭武，戴礼荣. 基于深层神经网络的藏语识别[J], 人工智能. 2015(3)
[69]余炜，周娅，万代立，杨喜敬. 基于改进 DTW 的彝语孤立词识别研究[J]. 昆明理工大学学报(自然科学版). 2014(5)
[70]张令通. 基于 HTK 的白族语音识别方法[J]. 大理学院学报. 2013(10)
[71]陈江，杨鉴，徐永华. 基于发音词典自适应的纳西语口音汉语普通话语音识别[C]. Proceedings of the 29th Chinese Control Conference(2013)
[72]Xi Ma, Xiaoxi Wang, Dong Wang, "Recognize Foreign New Words With Similar Words", Interspeech 2015
[73]Chao Xing, Dong Wang, Chao Liu, Yiye Lin, "Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation", NAACL 2015
[74]Gonzalez-Dominguez J, Lopez-Moreno I, Sak H, et al. Automatic language identification using long short-term memory recurrent neural networks[C]//INTERSPEECH. 2014: 2155-2159.
[75]Dehak N, Torres-Carrasquillo P A, Reynolds D A, et al. Language Recognition via i-vectors and Dimensionality Reduction[C]//INTERSPEECH. 2011: 857-860.
[76]Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//INTERSPEECH. 2010, 2: 3.
[77]Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 8609-8613.
[78]Deng L, Seltzer M L, Yu D, et al. Binary coding of speech spectrograms using a deep auto-encoder[C]//Interspeech. 2010: 1692-1695.
[79]Sainath T N, Kingsbury B, Ramabhadran B. Auto-encoder bottleneck features using deep belief networks[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 4153-4156.
[80]Kingma D P, Mohamed S, Rezende D J, et al. Semi-supervised learning with deep generative models[C]//Advances in Neural Information Processing Systems. 2014: 3581-3589.
[81]Shi Yin, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, Yinguo Li, "Noisy Training for Deep Neural Networks in Speech Recognition", EURASIP ASM 2015(2)
[82]Bourlard H A, Morgan N. Connectionist speech recognition: a hybrid approach[M]. Springer Science & Business Media, 2012.
[83]Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 4277-4280.
[84]Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 6645-6649.
[85]Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]//Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014: 1764-1772.
[86]Zhiyuan Tang, Dong Wang, "Recurrent Neural Network Training with Dark Knowledge Transfer", ICASSP 2016
[87]单煜翔,邓妍，刘加，一种联合语种识别的新型大词汇量连续语音识别算法，自动化学报，第 38 卷 第 3 期, 2012 年 3 月
[88]Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
[89]Povey D, Zhang X, Khudanpur S. Parallel training of deep neural networks with natural gradient and parameter averaging[J]. CoRR, vol. abs/1410.7455, 2014.
[90]Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.
[91]K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP 2014.
[92]S. Jean, K. Cho, R. Memisevic, Y. Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. 2014.

2. 项目的研究内容、研究目标，以及拟解决的关键科学问题（此部分为重点阐述内容）；

2.1　多少数民族语言语音识别研究定义

我们定义“多少数民族语言语音识别研究”为“**利用多个少数民族语言的数据和研究资源，结合少数民族语音和语言特色，研究复杂背景下的少数民族语音识别方法**”。这一研究可从建模和识别两方面进行分类。从建模上，我们将研究分为“**单语言建模**”和“**多语言建模**”，前者利用多个少数民族语言资源对某一少数民族语言进行增强建模，一般采用转移学习方法；后者利用多个少数民族语言资源对多个语言联合建模，一般采用共享学习方法。从识别上，我们将研究分为“**单语言识别**”和“**多语言识别**”两种，前者只支持单语言输入，后者支持多语言输入。

在实际研究过程中，需要从上述方法中选择一种建模方式和一种识别方式组成识别系统。不论选用哪种方式，我们都可称之为“多少数民族语言语音识别研究”，因为其中都用到了多少数民族语言的数据资源，用到了转移学习或共享学习方法，因此可以认为是一种多语言研究。当前基于 DNN 的多语言小语种研究多集中在“**多语言建模+单语言识别**”方式，即通过多语言数据提高单语言识别系统的性能；国内一些多少数民族语言识别系统（一般是汉语加一种少数民族语言）多数是“**单语言建模+多语言识别**”方式，即先对语言进行识别，再选择相应的识别系统进行解码，或多个单语言识别系统同步解码，再选择一个系统的输出作为识别结果。虽然不涉及混合语言问题，这些研究都属于学术界认可的“多少数民族语音识别研究”范畴。本项目基于这一范畴定义我们的研究内容，研究方向涉及各种建模和识别方式，但将重点放在“多**语言建模+多语言识别**”这一混合语言语音识别系统上。

2.2 研究内容

前一节我们分析了当前国内外多少数民族语言研究的现状，提出当前研究在建模、解码、语言覆盖性三方面的不足。为解决上述问题，我们三个方面展开研究：（1）将当前声学层多语言共享建模方法推广对词表和语言层的共享学习，解决当前建模方法中数据共享的局限性；（2）研究混合语言解码方法，解决当前解码方法中的延迟和计算开销问题；（3）研究有限资源建模方法，解决对资源极度稀疏的少数民族语言进行建模和识别的方法。表 1 总结了这三方面研究要解决的问题，当前方法和本项目所提出的方法。下面对这些研究方向做较详细的展开说明。

表 1：研究内容及主要方法

研究内容	问题来源	问题领域	当前方法	所提方法
多层次共享学习	语言共享	建模	DNN声学特征共享	声学、语言模型共享学习
多语言混合解码	语言混合	解码	语言识别加单语言解码	语言特征向量加多语言解码
资源稀缺语言建模	语言迁移	模型适应	很少	半监督学习

（一） 多层次共享学习研究

多语言语音识别研究的基本思路是利用多语言数据进行共享学习，以解决小语种数据资源稀缺问题，这一方法也是我们从事多少数民族语言语音识别的基本原则。然而，当前研究，无论是共享音素集(IPA 方法)还是共享发音特征（DNN 方法），都集中在声学层。本项目将声学模型共享学习扩展到词表和语言模型的共享学习。

（1）声学模型共享建模

声学模型共享建模是目前多语言小语种语音识别的通用方法，特别是基于 DNN 共享发音特征的方法是当前研究的前沿。其基本思想如图 1 所示：语言 L1 和 L2 共享从 Fbank 开始的特征提取过程(三层 DNN 网络)，提取的特征再用来对两种语言分别建立声学模型（两层 DNN)。

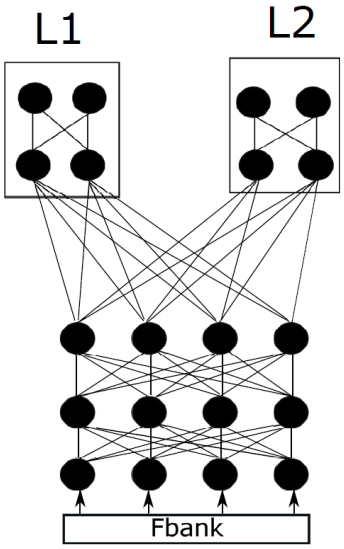

图 1 ：基于 DNN 的声学特征共享学习

本项目将在这一方面做进一步研究，特别是结合我国少数民族语言的特点进行扩展。具体包含如下三方面内容:

第一，我们研究语音识别领域的新技术对少数民族语言声学建模的作用，包括基于 Dropout[77]和 Noisy training 的网络正规化方法[81]，基于单状态 HMM+LSTM+MMI 的端对端训练方法(Chain 模型)，基于 Dark-knoledge 的模型转移学习方法等[86]。这些方法从原理上适合小数据和小模型训练，因而可用来提高少数民族语言声学建模能力。

第二，我们研究机器学习领域最近提出的一些方法对多少数民族语言声学建模的作用，特别是半监督学习方法和基于产生式模型的深度学习方法[65]。前者可以结合有监督学习和无监督学习，充分利用非标注数据，在小数据条件下提高模型性能； 后者利用产生式模型引入先验知识，基于少量数据完成模型自适应。这些方法在语音识别领域研究还不多。

第三，我们研究将多语言的共性和每种语言的特性相结合的差异性联合建模方法。这一方法是对半监督学习方法的扩展，通过引入每种语言的特性作为约束条件加入到 DNN 建模中，以实现对少数民族语言共性和特性的统一学习。

（2）语言模型共享建模

少数民族语言在语言层具有如下特性：（1）有限的词表资源；（2）有限的文本数据；（3）混合汉语、方言、网络用语的复杂表达；（4）各异化的词法、句法；（5）语言资源不均衡。

上述特性给少数民族语言模型建模带来很大困难。如果对每种语言单独建模，则数据不足，如果对多种语言建立混合模型，各异的词法、句法又使得多语言数据难以共享。

受声学建模中基于“语音特征”进行共享建模的启发，我们提出基于“语义特征”进行语言模型共享学习的方法。在声学建模中，共享学习基于不同语言在发音模式上的相似性；在语言建模中，我们认为不同语言在语义表达上也具有同样的相似性。具体而言，假设每种语言可以认为是由带有明确语义的实体词(E)和具有辅助功能的虚词(V)通过语法结构（L）组成，则虚词和语法结构是每种语言特有的，实体词可以在语义上跨语言共享。如图 2 所示：

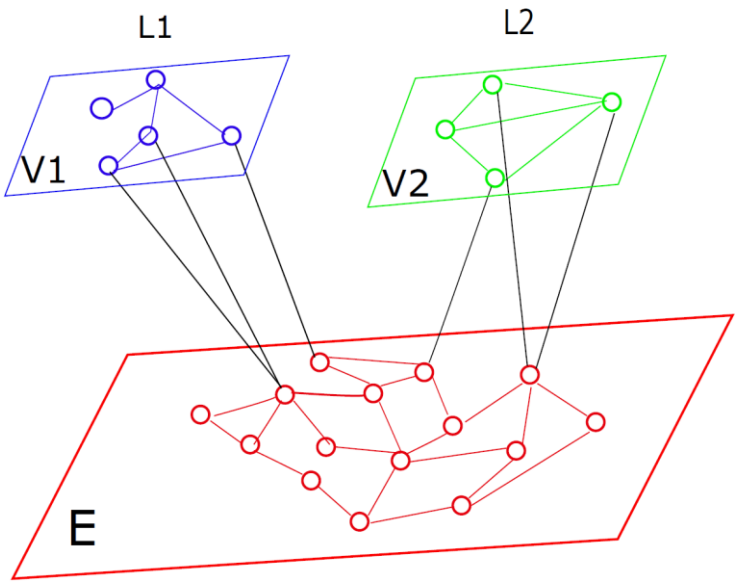

图 2：语言模型共享学习

在图 2 所示共享框架中，语言 L1 和 L2 具有不同的语法结构，拥有自己独立的虚词空间 V1 和 V2，但实体词共享一个语义空间，语言建模时对不同语言建立独立的模型，但共享实体词和实体词之间的连接概率。

我们特别关注基于 RNN 的语言模型（RNNLM）[76]。RNNLM 的基本思路是将字或词这些离散符号映射成低维空间的连续向量，以连续空间的转移概率代替 n-gram 模型中离散的概率表，从而对历史状态形成更强的共享。RNNLM 适合小语种多语言建模，一方面，基于 RNN 的参数共享机制， RNNLM 在小数据集上优势

更为明显；另一方面，基于 RNNLM 可能实现除实体词外虚词和语法结构的参数共享。我们将在“研究方法”一节中阐述可能的实现方法。

（二）多语言混合解码方法研究

如前所述，当前研究者从事的“多语言语音识别”多数是“多语言建模+单语言识别”，即通过多语言联合建模提高某一种语言的识别性能，而非多语言混合识别。当前支持多语言识别的系统一般采用“单语言建模+多语言识别”的模式，模型本身只支持单一语言，为实现多语言解码，需要一个语言识别模块，在解码前选择正确的语言进行单语言解码（图 3 左），或先进行多种语言的并行独立解码，在解码后选择合适的解码结果（图 3 右）。当前大多数多少数民族语言识别系统都采用这种方案。

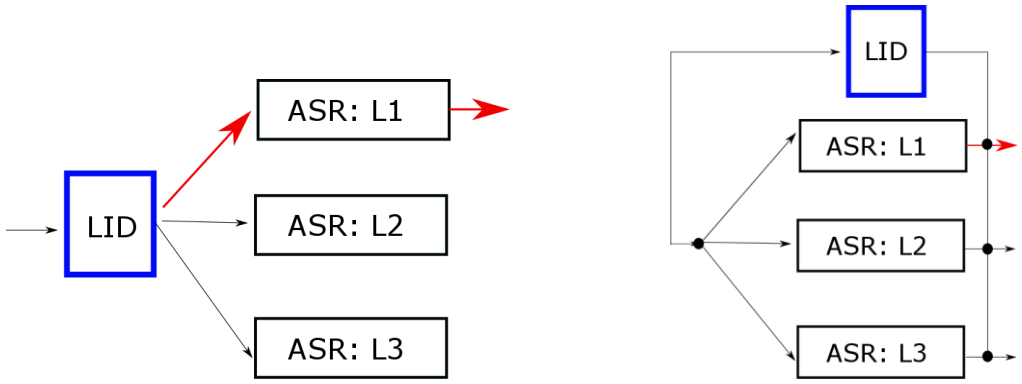

图 3. 基于语言识别(LID)的多语言解码

之所以采用这一方法，是因为多语言混合在一起会造成搜索空间的急剧膨胀，不仅降低搜索效率，而且会使解码的混淆度增加，导至识别率下降。通过语言识别检测，事实上是将多语言识别转化成单语言识别，避免了混合解码。然而，这一方法或者延迟显著（先做语言判断再做单语言识别），或者增加计算开销（先做多语言同步识别后做语言判断）。为减小延迟，有些研究在识别过程中对短语音片断进行语种识别[87]，这一方法并不改变单语言解码的本质。

本项目研究基于“多语言建模+多语言解码”的多少数民族语音识别方法，即在包含多语言词表的解码空间进行混合语言解码，直接找到最佳路径。该方法的特点是：（1）不需要语言识别，因而没有延迟；不需要多语言并行解码，因而计算开销较低；（2）多语言竞争，解码的结果同时包含语言标注和发音内容；（3）允许同一句话中语言的自由转换，如“我想说 I can fly”。

这一混合语言解码方法面临搜索空间膨胀和混淆度增加的困难。我们将研究利用语言向量对搜索空间进行约束的方法来解决这一问题。图 4 给出了该方法的基本思路：在识别过程中加入一个连续低维向量代表语言信息，这一向量在识别过程中对语言的表征越来越明确，因而对搜索空间的约束越来越强，逐渐将混合语言识别问题约束成单语言识别问题。这一思路可以认为是“单语言建模+多语言识别”方法的扩展，不过是将一次性语言识别替换成逐渐求精的语言向量，因而可以加入到混合解码过程中对搜索路径形成约束。我们将在“研究方法”一节具体阐述实现这一思路的方法。

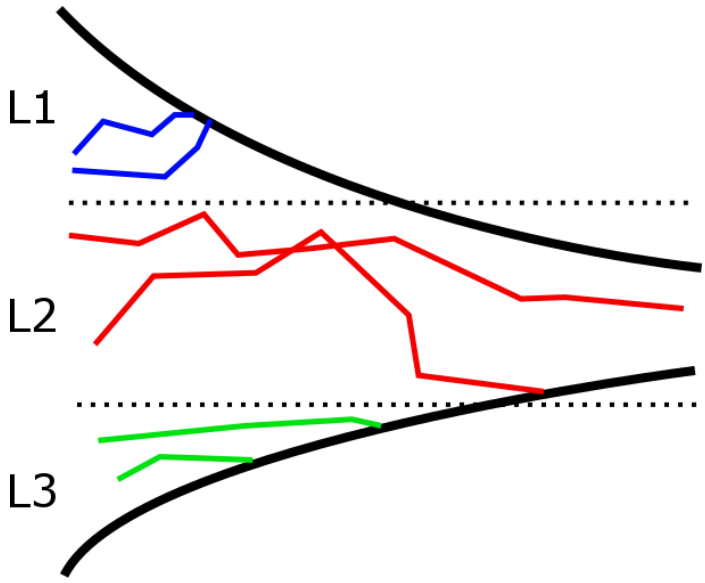

图 4：包含语言信息的混合语言解码方法

（三） 资源稀缺语言建模方法研究

当前少数民族语言语音识别研究基本集中在几个资源较丰富的语言上；对那些占绝大多数的资源极度稀缺语言，语音识别的研究还很少见到。本项目将对此展开研究。

我们说的“资源极度稀缺”是指如下几种情况：

a) 缺少语言模型训练数据，无法训练语言模型
b) 缺少语音模型训练数据，无法训练语音模型
c) 缺少语言学研究，无法得到词表
d) 缺少语音学研究，无法得到音素集和音素类
e) 缺少文字，无法形成书面表达

以上任何一种情况下，语音识别系统建模都会变得十分困难。基于我国少数民族语言的现状，本研究特别关注如下典型状况：缺少文字，缺少语音学研究，缺少语言学研究，无词表、无文本数据，唯一可用的是少量无标注的语音数据。对这些语言，语音识别十分困难，甚至定义都很模糊，因为这些语言本身就没有文字，“语音转文字”这一任务也就无从谈起。然而，这些语言覆盖我国众多少数民族地区，代表着我国文化的多样性，为它们提供信息化处理方案十分必要。

基于缺少文字这一特殊性，我们考虑用非监督学习方法将一种近似语言的识别系统自适应成资源稀缺语言的声学模型，再利用该模型进行音素标注。我们将研究选择近似语言的不同方法，研究多语言协同标注方法。

2.3 研究目标

a) **少数民族语音识别：**开发包括维、蒙、藏、哈、柯这五种语言的语音识别系统原型。通过共享学孔道，在这五种语言的种子数据上（无噪声）识别率达 85%以上，在实用数据上（含噪声及口语）识别率达 75%以上。
b) **多少数民族语言语音识别系统：**开发包括维、蒙、藏、哈、柯这五种语言的混合语言识别系统，实现实现混合识别方法下错误率降低在 5%以内，计算开销不随语言数的增长而增长，计算延迟增加在 20%以内。
c) **资源稀缺语言语音识别系统：**研究对资源极度稀缺语言的识别方法，实现音素标注的语音识别系统，音素标注正确率在种子数据上在 60%以上。
d) **资源建设：**本项目将积累面向少数民族语言的语音和语言数据库，并基于此设计语音识别基线系统。公开所收集的数据库、基线系统、标准测试集、工具，推动少数民族语言相关研究。

2.4 拟解决的关键科学问题

本项目解决的关键问题围绕“多语言”和“少数民族语言”这两点立项目标展开。这两方面研究既有独立性，又互相关联。多语言问题的困难主要在于解码时的混淆度，而少数民族语言问题的困难主要在于资源稀缺。这方面目标结合在一起，一方面使得识别任务更加困难，另一方面，多语言方法又为少数民族语

言识别提供了一种有效的解决方法，即共享建模。如何在共享建模时既利用多语言数据的优势，又减少多语言带来的混淆度，是所有问题的核心。同时，对那些资源极度稀缺的少数民族语言在共享建模基础上进行模型迁移，是将研究推广到绝大多数少数民族语言上时要解决的关键问题。我们将本项目要解决的关键科学问题归纳为如下四个方面：

1）多少数民族语言的协同建模

多语言协同建模是克服少数民族语言数据稀缺的有效方法。然而，我国少数民族语言众多，跨多个语系，不同语言之间差异性显著。如何对这些多样化、差异性的少数民族语言进行协同建模，具有很强的挑战性。

为此，我们需要研究如何在声学、词表、语言等多层进行共享建模，研究如何充分利用非标注数据；研究如何将多语言的共性和每种语言的特性相结合。这些都是需要在项目研究中需要回答的问题。

2）多少数民族语言的混合解码

当前多语言识别还是基于单语言识别系统（有可能通过多语言共享学习进行训练），或者延迟较大，或者计算量较大。多语言混合解码问题是本项目要重点解决的一个问题，在不降低识别性能的前提下，实现混合语言语音识别。

3）资源极度稀缺语言的处理方法。

我国大多数少数民族语言处于资源极度稀缺状态，不仅缺少语音和文本数据库，甚至连文字都没有。对这类语言构造语音识别系统显然异常困难。本项目将探索一系列方法，包括利用其它语言的书写系统为这些语言实现简单的音素标注和翻译。

3. 拟采取的研究方案及可行性分析（包括研究方法、技术路线、实验手段、关键技术等说明）；

3.1 研究方法与步骤

我们计划从如下四个方向展开研究：（一）设计少数民族语言的语音和语言数据库收集整理规范，并依此规范录制和标注五种少数民族语言的数据库，建立多语言语音识别的基线系统，开放相应的平台和工具；（二）基于所收集的多语

言数据库，展开多少数民族语言的共享建模研究；（三）研究多语言混合解码方法并完成多语言识别系统原型；（四）研究资源极度稀缺的少数民族语言语音识别方法。图 5 总结了这四方面研究包含的主要内容，我们将在下面详细展开说明。

<table>
<tr><td>数据库标准制定</td><td>工具建设</td><td>平台建设</td><td>数据公开与推广</td></tr>
<tr><td rowspan="3">数据采集</td><td>AM：基线系统优化</td><td>LM：n-gram 共享</td><td>语言识别+单语言解码</td></tr>
<tr><td>AM：半监督学习</td><td>LM：RNN 共享</td><td>RNN 语言-语音多任务学习</td></tr>
<tr><td>AM：差异性联合建模</td><td></td><td>多语言混合解码</td></tr>
<tr><td>基线系统</td><td>无监督自适应</td><td colspan="2">多语言协同标注</td></tr>
</table>

资源建设	共享学习	混合解码	资源稀缺语言建模

图 5：研究内容与方法

3.1.1. 数据、标准、工具、平台建设

本项目的一个重要任务是建立一套用于少数民族语音和语言研究的资源，包括数据、标准、工具和平台，以推动我国少数民族语言语音识别技术及相关研究。我们将采取如下步骤展开这方面的研究。

少数民族多语言数据库标准设计：尽管国内有不少研究者从事少数民族语言研究，也有相当研究单位构建了自己的数据库资源。然而，这些数据资源大多没有公开，因此影响力有限，标准化不足。目前为止，在语音识别方面，基本没有成规模的免费数据库公开出来，因此也缺乏公认的数据库设计和收集整理标准。欧盟 Babel 项目和 ESPRIT SAM 项目是国际上面向小语种的多语言数据建设项目，可以作为我们进行数据资源建设的参考。本研究将参考这两个项目的数据建设方法，设计我们自己的少数民族语音数据库收集整理标准。我们将考虑如下原则：

1）**任务适用性原则：**保证收集的数据集对我们的核心研究任务（语音识别）具有直接适用性，同时兼顾多语言语音学分析、说话人识别、语种识别等相关任务的需求。

2）**种子数据核心研究相关原则：**少数民族语言数据收集和标注的成本高，因此必须保证种子数据对核心任务的针对性和有效性。为满足这一要求，我们对每种语言设计一个种子数据集(seed set)，该数据集规模较小，在录制时排除噪

音、信道等干扰，突出语言上的差异性；同时，语音内容在录制前经过认真设计，保证音素和说话人上的广泛覆盖性。

3）实用性原则：由于种子数据只关注核心任务上的区分性（如发音内容）而忽略其它干扰因素，如噪音信道等，这意味着由该数据训练的模型与实际应用环境不匹配。为保证我们研究的结论尽可能符合实际应用环境，我们将在种子数据集基础上收集和标注主体数据库(body set)。该数据库内容并非事先设定，而是来源于实际录音数据的标注。

少数民族语言的语音数据库采集： 依上述数据库收集整理规范，本项目将收集维吾尔语、蒙语、藏语、哈萨克语、柯尔克孜语等五种少数民族语言，其中维吾尔语包括三种方言：中心、和田、罗布方言，三个方言的主要差别表现在语音上。标准语以中心方言为基础，占总人口 80%，和田方言人口大约占 19%，罗布方言人口占 0.5%，并在逐渐减少。除此之外各个地方都有次方言，如：喀什方言，吐鲁番方言等。藏语各包括三种方言:卫藏、安多和康巴方言。其中卫藏方言以标准拉萨话为基准，安多方言以夏河话为基准，康方言以德格话为基准。中国境内的蒙古语分为三个方言：内蒙古方言（也称中部方言主）、巴尔虎布里亚特方言(也称东北部方言)和卫拉特方言（也称西部方言），其中内蒙古方言分布最广，使用人口也占中国蒙古族人口的 90%以上，所以本项目只考虑内蒙古方言。

对维吾尔、蒙、藏、哈萨克这四种覆盖人口较多的少数民族语言，我们计划采集较多的语音数据。对 seed set，我们设计语音文本以保证音节和音素覆盖，每种语言通过电话录音方式采集 50 小时 100 人的电话录音；对于 body set，我们不预制文本，每种语言通过模拟电话客服方式采集自由对话文本，得到约 200 小时 300 人的电话录音。录音采样率为 8k，采样大小为 16bit。seed set 的录音设备不变，body set 的录音设备自备。

柯尔克孜语覆盖的人口较少，我们计划采集相对较少的语音数据。对 seed set，我们计划采集 20 小时 20 人的电话录音，对 body set，我们采集 50 小时 50 人的电话录音。

少数民族语言的文本数据库采集：对每种语言，我们将收集整理 50M 文本数据库，用以训练语言模型。

其它资源和工具整理：对每种语言，整理音素集、音素类、词表、与其它语言互相翻译的词典等资源。收集、整理、设计和实现一系列用于语音和语言处理

的工具，包括分词工具，词干提取工具，音节分析工具等。

基线系统构造：基于所收集到的五种少数民族语言的语音数据库和其它资源，参考 Kaldi 的 THCHS-30 汉语语音识别系统构造方法，构造基于 DNN 的语音识别基线系统。

评测与推广：本项目将利用所采集到的数据作为标准测试数据集，组织若干面向少数民族小语种的语音识别和说话人识别技术评测，进一步推动少数民族小语种语音处理技术的研究。

平台建设与数据公开：我们将开放一个少数民族语音语言研究合作平台，本项目所采集到的所有语音、文本数据、语言资源、基线系统构造方法将在该平台上全部公开，以研究为目的使用将完全免费。该平台还将提供对少数民族语言的众包标注功能，吸引研究者为少数民族语音语言研究做贡献。

3.1.2 多少数民族语言共享建模研究

当前多语言小语种识别研究集中在基于 DNN 的声学特征共享。本研究将从这一基本方法开始，推广到词表、语言模型等多层次的共享学习。

1）声学模型：面向少数民族语言的基础模型优化

本研究首先将最近提出的几种新型建模方法和优化方法移植到少数民族语言声学建模中。其中建模方法包括：作为特征提取的 TDNN 模型[82]，CNN 模型[83]，作为动态建模的 RNN/LSTM 模型[84]，作为训练目标的 CTC 准则[85]；优化方法包括：基于 Dropout 的模型鲁棒学习方法[77]，基于 Recifier 非线性函数的稀疏化方法[77]，基于 Batch Norm[88]和 Natural Gradient[89]的准二阶优化方法，基于 Noisy traning 的抗噪声方法[81]。上述这些方法在英语和汉语的若干识别任务中取得了很好的效果，其中有些方法特别适合数据稀缺条件下的声学建模，如 Dropout, noisy training 等。我们将测试这些方法在少数民族语言上的建模效果，以得到一个较强的基线系统。

2）声学模型：多语言训练与转移训练方法

上述基础模型的优化方法适合单一语言建模。为进一步提高建模能力，克服少数民族语言资源稀缺的困难，我们采用多语言训练与转移训练方法。多语言训练采用 Hybrid 或 Tandem 结构，将若干少数民族语言（可包括汉语）的语音数据放在一起训练，共享特征提取层。这一方法已经被广泛应用在小语言声学模型建模中。一个需要探索的问题是上述新的模型方法（如 LSTM, CTC）对多语言训练

是否兼容，是否可以取得好的效果。另外，如何合理调节多语言在模型训练中的权重，在充分利用数据资源的同时平衡不同语言在模型中的代表性，也是需要探索的问题。

转移学习是另一种多语言声学模型训练方法。本研究中，我们将利用汉语积累的丰富语料对少数民族语言进行转移学习。典型的方法是利用汉语的语料进行预训练(pre-training)，保留 DNN/RNN 模型的低层作为初始化参数训练少数民族语音模型。我们前期实验表明利用中文对维吾尔语模型进行预训练可以提高维吾尔语 DNN 声学模型的性能。然而，如何更有效地利用汉语（或其它资源丰富语言）的语音数据还需要进一步探索。本项目将研究一系列新方法。一种方法是通过引入一定的正规化因子对转移训练中的自适应进行规范，如最近提出的基于 KL 距离的正规化方法[54]。我们将尝试更有效的正规化方法，利如分层正规化方法，即越底层网络正规化越强，越上层网络正规化越弱，以保证底层语音特征的稳定。另外，基于 KL 距离的正规化方法可以认为是贝叶斯方法的一种，即对 DNN 输出加入一个先验概率。这一方法可以扩展到对 DNN 的隐藏层输出或 DNN 参数加入先验概率，即基于贝叶斯的 DNN 自适应学习方法。这一扩展不仅可以对网络进行更灵活的自适应，还可以扩展到基于其它训练准则的训练过程中，如 CTC 和 MPE。我们将研究这些方法在转移训练中的效果，并探究这些方法在不同语言上的有效性，研究这些方法在多语言转移训练（学习目标为多语言 DNN）中的有效性。

3）声学模型:差异性联合建模

上述多语言共享学习多关注多语言的共性，对每种语言的特性并没有考虑，因而可能带来建模上的偏差。一个典型例子是在中文和维吾尔语共享训练中，如果我们不考虑维吾尔语和中文在数据量上的差别，生成的模型将强烈偏向中文。

本研究提出将语言共性和特性结合起来的共享学习方法。基本思路是，在保持总体共享学习目标不变的前提下，将各个语言的特性作为训练的约束条件，加入到目标函数中。该方法可表示为：

$$L(D;\theta) = \sum_i L(D_i;\theta) + \sum_i\sum_j k_{i,j}(D_i;\theta)$$

其中 $L(D;\theta)$为约束后的目标函数，D 为全部训练数据，θ 为模型参数。i 代表某种语言，D_i 代表语言 i 的全部数据， $L(D_i;\theta)$为语言 i 的原始训练目标（如 phone state cross entropy），$k_{i,j}(D_i;\theta)$为在语言 i 上的第 j 个约束条件。图 6 给出这一方

法的说明。在原有训练准则(obj0)的基础上，加入一系列约束条件，最后的训练目标包括原有训练目标和加入的约束。

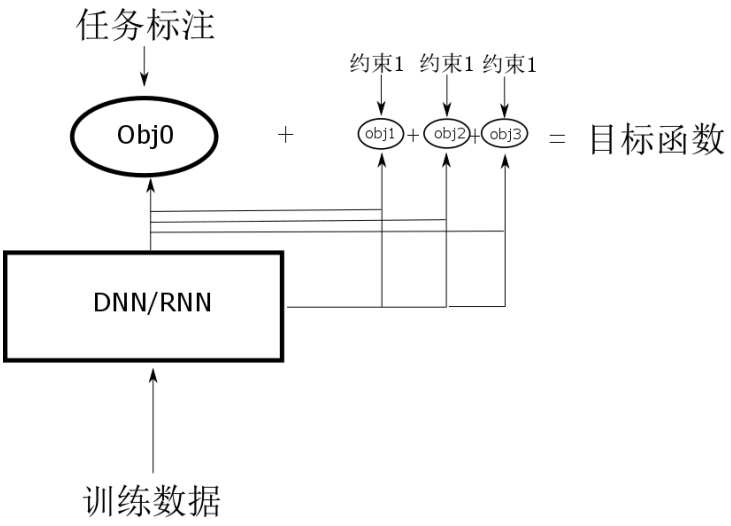

图 6. 差异化多语言声学建模

在前面中文和维吾尔语共享训练的例子中，我们可以在训练目标中加入一个数据量约束项，使模型训练更平衡。事实上，我们在少数民族语言建模中所遇到的问题几乎都可以归结为约束问题，进而用上述方法解决。

4)声学模型:非监督和半监督学习

上述差异化建模的一个典型例子是无监督和半监督学习[80]。少数民族语言数据稀缺，但非标注数据相对容易得到。利用无监督学习或半监督学习方法，可以充分利用这些非标注数据，提高建模质量。一种典型的方法是在标准 DNN 建模的基础上加入一个自编码模型(Autoencoder, AE)[78][79]，如图 7 所示，其中 DNN 的部分输出结果经过另一个 DNN 后，恢复成原训练数据。

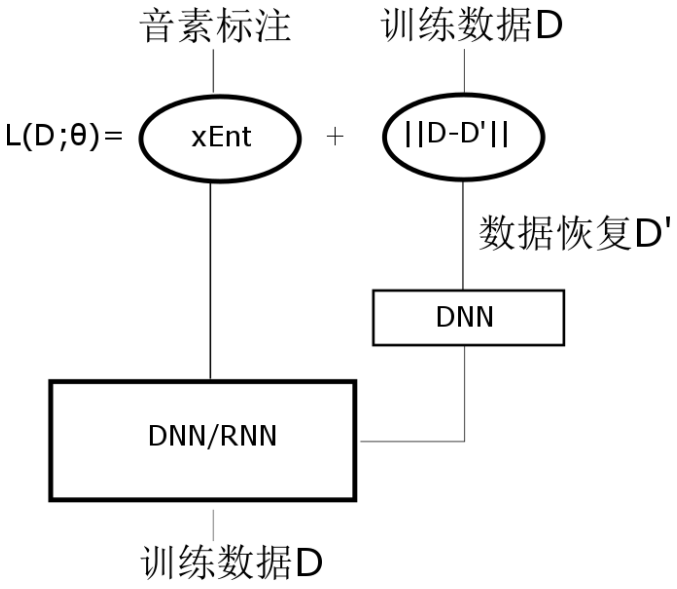

图 7. 半监督声学建模

加入约束后的训练准则包括两部分：对音素标注的预测误差和对训练数据的恢复误差，表示为：

$$L(D;\theta) = \sum_i L(D_i; \theta) + \sum_i ||D_i - D_i'||$$

其中 D_i' 表示对第 i 种语言的训练数据经过另一个 DNN 后被恢复的数据，||.|| 表示二阶规范。这一方法可以扩展到其它无监督学习模型，特别是产生式无监督模型（如 GMM 或 Gaussian RBM）。

5)语言模型：n-gram 多语言共享模型

当前对多语言语音识别的研究多集中在声学模型上，事实上语言模型面临更多挑战。一方面，多语言问题本质上是语言层的差异性问题，另一方面，少数民族语言的文本数据同样稀缺。借鉴声学建模中共享学习思路，本研究将探索利用共享学习提高语言模型建模能力的方法。我们首先讨论基于 n-gram 语言模型的共享学习方法，下节讨论基于 RNNLM 的共享学习方法。

对于 n-gram 语言模型，词间概率由一个离散的概率分布表表示。这些概率信息描述一个词序列产生下一个词的概率，等价于一个有限状态自动机(FST)，即语音识别中的解码图。

我们假设语言模型包括实体词 E，虚词 V 和语法结构 L。以 bi-gram 模型为例，E 和 V 相当于解码图上的结点，L 相当于解码图上边的权重。我们假设实体词和实体词之间的权重 L(E)可以在语义上跨语言共享，而虚词 V 及虚词之间的权重 L(V,V)和虚词与实体词的权重 L(V,E)不可共享。这一共享结构如 2.2.1 节中的图 2 所示。

上述共享模型需要将不同语言中表达相同语义的实体词映射到一个共同的语义符号上，再集合多语言的训练语料训练多语言模型，实际上生成了如图 2 所示的混合语言解码图。我们在前期工作中利用这一模型解决汉语中混杂英语单词的问题，取得了很好的效果[72]。

语义映射是上述模型的关键。我们考虑如下几种映射方法：

a)词典映射：用双语言翻译词典直接得到同义词映射。

b)多语言语义网映射：多语言语义网包含同一实体的不同语言表达。本项目将以汉语为基础，收集一系列多语言实体列表及其关系，并开放在线工具，鼓励研究者对此列表和关系进行补充。

c)跨语言词向量空间映射：用词向量工具（如 word2vect）对每种语言训练词向量，再利用少量词典映射得到跨语言词向量空间的线性映射。我们前期工作表明这种线性映射可以将英语和法语嵌入到同一语义空间[73]。

6)语言模型：RNNLM 多语言共享模型

基于深度学习的语言模型，特别是 RNNLM 近年来得到广泛关注[76]。实验证明，RNNLM 在小规模数据集上可以取得比传统 n-gram 更低的复杂度和更好的识别率，因此适合少数民族语言模型建模。

更重要的是，RNNLM 有可能提供一种更有效的多语言共享学习方法，也更接近声学模型共享学习方式。如图 8 所示，其中 L1 和 L2 是两种语言，红色结点代表共享的语义符号（跨语言实体词集合），蓝色结点代表 L1 独有的虚词，绿色结点代表 L2 独有的虚词，黑色结点代表隐藏层的 LSTM 单元。

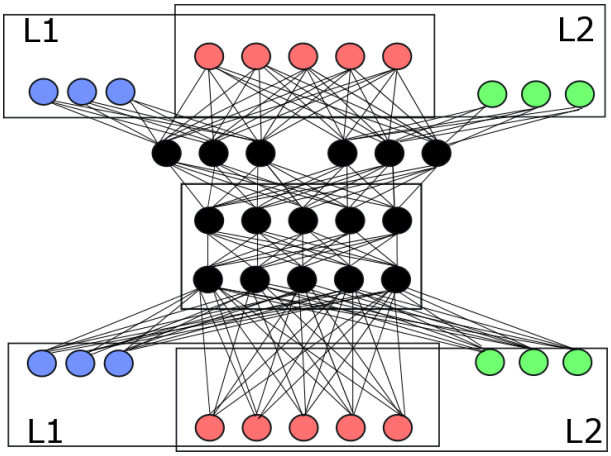

图 8：基于 RNN 的语言模型共享学习

比较图 1 的声学模型共享方式， 我们看到图 8 的语言模型共享非常类似，都是共享中间隐藏层。在声学建模中，这部分共享结构用来提取共有的声学特征，在语言建模中，这部分不仅用来提取共有的语义特征，同时也提供语法结构的参数共享。

3.1.3 多语言解码方法

1）多语言混合解码框架

多语言识别方法分为如下两种方案：独立语言解码和混合语言解码。在独立语言解码中，虽然语音模型和语言模型都可能支持多语言，但搜索空间被限制在某一种语言中，因此实际上是单语言解码。为实现多语言语音识别，一般需要一个语言识别模块，在解码前选择正确的语言进行解码，或先进行多种语言并行独立解码，在解码后选择某一种语言的解码结果作为输出。参考 2.2.1 节图 3。当前系统大多基于这种独立解码方案。

独立语言解码或者延迟较大（前端语言判决），或者计算开销较大（后端语言判决）。特别是这一方法无法支持多语言共享语言模型：虽然语言模型和搜索空间实现了共享，但独立语言解码并不能充分利用这种共享优势。

理想的方法是用统一的多语言解码图（如图 2,4），在解码过程中自动选择最优语言和该语言上的最优路径，这一方法称为混合语言解码。然而，我们的前期实验发现，当解码图扩展到多语言时，系统的性能会显著下降。分析可知，混合解码的困难主要有两点：（1）当语言数较多时，混合解码的解码图急剧扩大，在有限的 beam 约束下几乎不太可能将最优结果保留下来；（2）当语言较多时，

不同语言在解码时的先验概率变得很难平衡，一句话很可能被识别成其它语言中具有类似发音的句子。

当前混合解码研究一般只局限在一种宿主语言中加入其它语言的少量词汇。这种情况只增加声学模型的状态空间，而语言模型状态空间基本保持不变，因此不会带来太大问题。最近，百度的 DeepSpeech2 提出用 End-to-End 方法可实现中文和英文的混合解码。这是因为该模型中只有两种语言，且这两种语言差距较大。特别是，这两种语言的数据都很丰富，因此 LSTM 可以同时学习语言信息和发音内容。如果加入更多语言，且都是小语种，依靠 End-to-End 模型实现混合解码依然很困难。

为了解决混合解码搜索空间膨胀带来的混淆度，本项目研究一种渐进式语言约束解码方法。我们主要思路包括两个：（1）语言信息追加 （2）平衡路径搜索。基本假设是,当最初开始解码时，我们并不知道是何种语言，因此所有语言的局部路径都会保留。我们可以放大 beam，保留更多路径，特别是对不同语言保留相近的路径数目，这意味着语言的先验概率均衡（也可以用不均衡的先验概率，如基于前一句的解码结果）。随着解码进行，语言信息逐渐累积，依据这些信息对语言先验概率进行修正，并用来约束下一时刻的解码。随着解码的进行，累积的语言先验概率越来越准确，解码将逐渐收缩到正确的语言上来。

形式化这一混合解码过程。首先确定混合解码事实上包含两个任务，(1)得到正确的语言 L，(2)得到该语言下正确的路径 W，即最大化下式:

$$P(W,L|O)=P(W|O,L)P(L|O)$$

其中 $P(W|O,L)$是单语言解码的目标函数，$P(L|O)$为给定发音后的语言后验概率（相应为解码路径的先验概率）。传统方法将计算 $P(L|O)$当作独立过程（语言识别），我们的方法将 $P(L|O)$作为解码过程的一部分，因而具有实时性。注意:

$$P(L|O_{1:t})\propto P(O_t|L)P(L|O_{1:t-1})$$

因此语言信息在解码的每一步可以逐渐累积更新，这意味语言信息会越来越准确，将解码过程逐渐约束到正确的语言上来。前面 2.2.1 节图 4 描述了通过语言信息积累，逐渐把解码约束到语言 L2 上的过程。刚开始时各语言的先验近似，每种语言都有部分路径保留下来。随着解码的进行，L2 的先验知识越来越强，其它语言的解码路径被陆续剪掉,仅有 L2 的路径保存到最后。

2 ）基于多任务 RNN 的语言概率计算方法

在上述多语言混合解码框架中，语言概率 P(L|O1:t)具有重要意义。一种简单方法是基于当前成熟的语言识别系统，如 i-vector 系统[75]，在每一时刻计算出一个语言向量，利用该向量计算 P(L|O1:t)。最近研究表明，利用 RNN/LSTM 可以实现语言识别[74]。与 i-vector 相比，这种方法不需要在每一时刻对 i-vector 进行计算，因此更简洁。特别是，我们发现这一语言概率计算方法和语音识别系统可以统一到一个 RNN 模型中，实现多任务学习。

所谓多任务学习，是指对多个相关任务同时学习，使每一任务通过共享信息都能得到性能提高。当前多语言声学模型共享建模即是一个典型的多任务学习。我们可以将语音识别和语言识别作为两个相关的任务，通过多任务学习共同提高两者的建模质量。特别是，如果语言识别可以在每一时刻为语音识别提供一个语言概率，则上节所述的混合解码变得十分自然。

一种可行的语音/语言多任务学习结构如图 9 所示，其中语音识别在某一时刻的输出作为语言识别在下一时刻的输入；同样，语言识别的输出作为语音识别下一时刻的输出。这实际上等价于一种分块多任务 RNN 结构，每一块 RNN 结构对应一个任务，并生成下一时刻对其它任务的输入。注意在图 9 中我们只表示了不同任务之间的反射层，事实上同一任务一般也会有反射层。例如在语言识别中，为累积语言信息，这种 RNN 结构通常是必要的。

图 9 所示的 RNN 结构将语音识别和语言识别这两个相关任务统一到一个完整的 RNN 模型中，简化了建模过程和训练流程。特别重要的是，这一网络同时完成这两个任务，一方面可以互相提高各自的识别性能，同时又会为对方及时提供互补信息（如 P(L|O1:t)），这一点在多语言混合解码时至关重要。

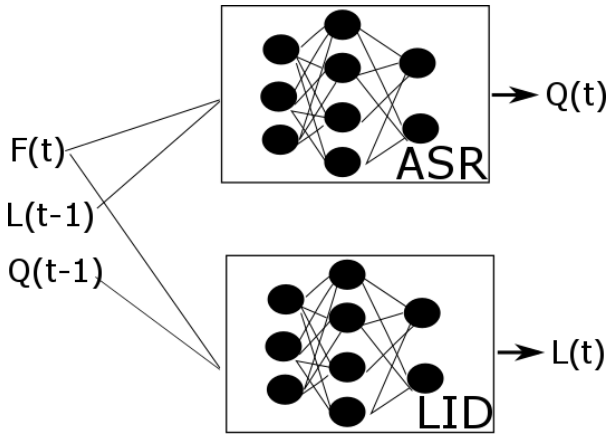

图 9：基于 RNN 的语音/语言多任务学习结构

3.1.4 资源极度稀缺的少数民族语言语音识别

我国少数民族语言众多，大多数少数民族语言处于资源极度稀缺状态。这不仅因为缺少人力进行语音和文本数据采集、整理、标注，更因为很多少数民族语言仅停留在口语阶段，缺少文字。同时，对这些语言的语音学和语言学研究也非常有限，不仅缺少较完整的词表，有很多连音素都没有整理出来。唯一可用的资源是一些没有标注的语音数据（因没有文字，没有音素集，所以无从标注）。对于这些语言的语音识别研究还相当有限。

因为缺少本民族文字和音素集，我们需要借用其它语言（借用语言）的音素集对资源稀缺语言（目标语言）进行建模。为寻找到最合适的借用语言，我们可以使用两种方法（1）声学模型最大似然法：在所有候选语言的声学模型上计算语音数据的似然值，选择最大似然的语言作为借用语言；（2）语言识别法：利用语言识别系统，计算与目标语言最相近的语言作为借用语言。

选定借用语言后，可以用半监督学习方法进行模型自适应。一个可能的流程如下：

（1）使用借用语言的声学模型对目标语言的训练语言做音素解码

（2）利用得到的解码结果对声学模型进行自适应

（3）利用得到的解码结果训练音素语言模型

（4）使用适应后的声学模型和训练得到的语言模型重新解码训练数据

经过几次迭代后，得到的声学模型和语言模型可用于对目标语言进行音素级识别。一个可能的改进方法是对识别出的音素串进行重复模式发现，归纳出常出现的模式，列为常用词，训练包含音素和词的混合语言模型。

另一个有可能的方法是利用当前五种语言对稀缺资源语言进行标注，再将这五种标注结合起来，形成统一标注。这一方法可防止一种语言进行标注产生的错误（如某些发音借用语言无法覆盖）。但是，这一方法需要一个公共音素表，需要计算每种语言标注的置信度，且计算开销比较大。

3.3 可行性分析

本项目提出的研究方法具有可行性。第一，深度学习技术在语音识别领域已有较多较究，在小语种多语言识别方面也取得了一系列研究成果，证明基于DNN/RNN 的少数民族多语言语音识别是可行的。第二，申请人及其团队在语音识别、少数民族语音处理方面具有丰富的经验，积累了众多资源和科研成果（见申

请人和项目组简介)。第三，我们已经在相关领域开展了一系列相关研究，一些研究和本申请所提出的方案直接相关（见 2.2 和 3.1)。现有实验结果表明，本项目提出的方案是可行的，且有些设想已经在汉语数据或小规模实验中得到验证。例如， 我们基于转移学习训练的维吾尔语识别系统已经上线[4]。

4. 本项目的特色与创新之处；

本项目的特色在于瞄准国际最新研究方向，利用清华大学语音语言研究中心的技术和人才优势，结合新疆大学和西北民族大学的资源优势，建设中国第一个少数民族语言的多语种开放数据资源和平台，为推动语音识别和相关领域的研究打下基础。同时，本项目充分吸收国际上在这一方向上的研究经验，并结合我们自己在机器学习、语音识别、小语种语音信息处理等领域里的优势，利用最新的机器学习方法解决多少数民族语音识别中的关键问题。我们的创新之处在于：

1) **基于深度学习的多层次共享训练方法，较之当前方法更适合多少数民族语言共享建模；**
2) **实现混合语言语音识别，支持多语言自由转换；**
3) **实现资源极度稀缺少数民族语言的音素标注。**

5. 年度研究计划及预期研究结果（包括拟组织的重要学术交流活动、国际合作与交流计划等)。

5.1 年度研究计划

本研究计划在五年内完成，分为五个工作模块:

W0: 项目管理：进度控制，代码管理，发表与发布
W1: 资源建设：数据、工具的积累与发布，共享平台建设
W2: 多语言共享学习：声学与语言模型多语言建模
W3: 多语言识别系统：混合语言解码研究，多少数民族多语言识别系统原型构建
W4: 资源稀缺语言建模：音素标注

项目研究进程 Gantt 图如表 2 所示。

[4] http://www.sinovoice.com/news/company/2015/0210/152.html

表 2. 项目研究计划 Gantt 图

	YEAR1			
	Q1	Q2	Q3	Q4
W0[项目管理]	项目启动:协同工作方式，调配计算资源		参加 Interspeech	研讨会 A；年终总结会 A
W1[资源建设]	整理现有资源，设定资源标准			发布数据资源 A
W2[多语言共享学习]		建设五种语言基线识别系统		
W3[多语言识别系统]	基于 i-vector 的 LID		基于 LID 的多语言识别系统原型	
W4[资源稀缺语言建模]			有文字的半监督自适应学习	
	YEAR2			
	Q1	Q2	Q3	Q4
W0[项目管理]	参加 ICASSP		参加 Interspeech	研讨会 B；年终总结会 B
W1[资源建设]	数据库收集			发布数据资源 B
W2[多语言共享学习]	声学模型转移学习；n-gram 转移学习		声学模型转移学习+共享学习	
W3[多语言识别系统]	DNN-LID 多语言识别系统		双语言 n-gram 混合解码	
W4[资源稀缺语言建模]	有文字的半监督自适应学习			
	YEAR3			
	Q1	Q2	Q3	Q4
W0[项目管理]	参加 ICASSP		参加 Interspeech	研讨会 C；年终总结会 C
W1[资源建设]	数据库收集与工具建设			发布数据资源 C
W2[多语言共享学习]	RNNNLM 转移学习		多语言 n-gram LM	
W3[多语言识别系统]	双语言 RNNLM 混合解码		多语言 RNNLM 混合解码	
W4[资源稀缺语言建模]	无文字无监督自适应学习			
	YEAR4			
	Q1	Q2	Q3	Q4
W0[项目管理]	参加 ICASSP		参加 Interspeech	研讨会 D；年终总结会 D
W1[资源建设]	数据库收集与工具建设			发布数据资源 D
W2[多语言共享学习]	多语言 RNNLM		非监督与半监督学习	
W3[多语言识别系统]	多语言 RNNLM 混合解码			
W4[资源稀缺语言建模]	多语言音素标注			
	YEAR5			
	Q1	Q2	Q3	Q4
W0[项目管理]	参加 ICASSP		参加 Interspeech	项目总结会
W1[数据整理收集]	数据库收集与平台建设			发布全部数据
W2[多语言共享学习]	差异性多语言建模			
W3[多语言识别系统]	结合多任务学习的多语言 RNNLM 混合解码		原型系统开发，发布	
W4[资源稀缺语言建模]	多语言音素标注			

5.2 其它学术进度及国际交流计划

2017/01~2017/12：参加国际学术会议 ICASSP、Interspeech。参加国内学术会议 NCMMSC, ChinaSip。邀请国际知名专家访问一次。发表 SCI 论文 2 篇，EI 论文 5 篇。提出发明专利申请 2 项。

2018/01~2018/12：参加国际学术会议 ICASSP、InterSpeech 及 NCMMSC、ChinaSIP。组织关于少数民族语言深度学习技术学术会议一次，组织。邀请国际知名专家访问一次，发表 SCI 5 论文两篇，EI 论文 5 篇。提出发明专利申请 2 项。

2019/01~2019/12：参加国际学术会议 ICASSP、InterSpeech 及 NCMMS、ChinaSIP。组织第二届关于少数民族语言深度学习的学术会议，邀请国际知名专家访问一次。发表 SCI 论文 2 篇。发表 EI 论文 5 篇。提出发明专利申请 2 项。

2020/01~2020/12：参加 ICASSP、Interspeech 及 NCMMSC、ChinaSIP。组织第三届关于少数民族语言深度学习的学术会议。邀请国际知名专家访问一次。发表 SCI 论文 2 篇。发表 EI 论文 5 篇。提出发明专利申请 2 项。

2021/01~2021/12：参加 ICASSP、Interspeech 及 NCMMSC、ChinaSIP。组织第四届关于少数民族语言深度学习的学术会议。邀请国际知名专家访问一次。发表 SCI 论文 2 篇。发表 EI 论文 5 篇。提出发明专利申请 2 项。

5.3 预期研究成果

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。

参与领域权威学术会议 Interspeech 和 ICASSP 不少于 10 次。

组织至少 4 次关于少数民族语言深度学习的学术沙龙。

邀请国际领域知名专家访问报告不少于 4 次。

（二）研究基础与工作条件

1. 研究基础（与本项目相关的研究工作积累和已取得的研究工作成绩）；

1.1 科研成果

申请人及其团队长期从事语音识别、自然语言理解和机器学习领域的研究，在国际一流期刊和会议发表过论文多篇。下面列出我们近两年内取得的和本项目直接相关的研究成果。这些成果是本项目中所列科研计划的预研。

1) 近两年来小语种语音处理研究成果 [SCI 标红]

1. Mijit Ablimit、Tatsuya Kawahara、Askar Hamdulla ，Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language, Speech Communication, pp 78-87, 2014.5

2. Askar Rouze, Shi Yin, Zhiyong Zhang, Dong Wang, Askar Humdulla, Fang Zheng, "THUYG THUYG-20: A free Uyghur Speech Database", Tsinghua Xuebao (in Chinese)

3. Askar Rouze, Dong Wang, Zhiyong Zhang, Thomas Fang Zheng, "Speaker Recognition with UYGTH-20", O-COCOSDA 2015

4. Rozi Askar, Dong Wang, Fanhu Bie, Jun Wang, Thomas Fang Zheng, "CROSS-LINGUAL SPEAKER VERIFICATION BASED ON LINEAR TRANSFORM", ChinaSIP 2015.

5. 艾斯卡尔•艾木都拉，赛尔达尔•雅力坤，祖丽皮亚•阿曼，地里木拉提•吐尔逊. 维吾尔语浊塞音的声学特征分析. 清华大学学报(自然科学版),2013,06:828-834.

6. Askar Hamdulla, Guljamal Mamateli, Askar Rozi. Hierarchical prosodic boundary prediction for Uyghur TTS. 4th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2012)

7. Mijit Ablimit, Tatsuya Kawahara, Askar Hamdulla. Discriminative approach to lexical entry selection for Automatic Speech Recognition of agglutinative language. In Proc. ICASSP 2012.

8. 胡阿旭，吕士良，格根塔娜，于洪志.蒙古语松紧元音的言语空气动力学研究[J]. 清华大学学报(自然版)(EI). 第 51 卷，第 9 期：1167-1170.

9. Hongzhi Yu, Gegentana, AxuHu. An Aerodynamic Study of Aspirated

and Unaspirated Consonants in Mongolian[C]. The 2nd International Conference on Mechanic Automation and Control Engineering (MACE 2011), Vol.1:1015-1018.

10. 李冠宇，孟猛. 藏语拉萨话大词表连续语音识别声学模型研究[J],计算机工程,2012,3(5)，189-191

通过上述工作，我们积累了一定量的维吾尔语语音数据，实现了基于深度学习的维吾尔语语音识别系统。我们发现基于 DNN 转移学习方法可以有效提高维吾尔语识别性能（识别率再提高 10%以上）。同时，我们发现考虑语言的特殊性可进一步提高识别率，如基于维吾尔语造词的粘着性，对 morphone 建模可提高识别率 3%以上。特别是工作（2）和（3），我们发布了 20 小时免费维吾尔语数据库，用于语音识别和说话人识别，在少数民族语言数据资源开放方面走在了前列。

2) 近两年来语音识别研究成果 [SCI 标红]

1. Meng Sun, Xiongwei Zhang, Hugo Van hamme, and Thomas Fang Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement", IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 93-104, Vol. 24, No. 1, Jan. 2016

2. Shi Yin, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, Yiguo Li, Noisy Training for Deep Neural Networks in Speech Recognition, EURASIP Journal on Audio, Speech, and Music Processing, 2015(2).

3. Javier Tejedor; Doroteo T. Toledano; Dong Wang; Simon King; Jose Colas, "Feature analysis for discriminative confidence estimation in Spoken Term Detection", Computer Speech and Language,28 (2014), pp. 1083-1114.

4. Dong Wang, Ravichander Vipperla, Nick Evans, Thomas Fang Zheng, "Online Non-negative Convolutive Pattern Learning for Speech Signals", IEEE trans. on signal processing, vol.61, no.1, pp.44-56.

5. Zhiyuan Tang, Dong Wang, Zhiyong Zhang, " Recurrent Neural Network Training with Dark Knowledge Transfer", ICASSP 2016

6. Mian Wang, Dong Wang, "VMF-SNE: EMBEDDING FOR SPHERICAL DATA", ICASSP 2016

7. Dong Wang, Thomas Fang Zheng, "Transfer learning for speech and language processing", APSIPA 2015, invited paper

8. Mengyuan Zhao, Dong Wang, Zhiyong Zhang, Xuewei Zhang, "Music Removal by Denoising Autoencoder in Speech Recognition", APSIPA 2015, best paper

9. Xi Ma, Xiaoxi Wang, Dong Wang, "Recognize Foreign Low-FrequencyWords with Similar Pairs", Interspeech 2015.

10. Xiangyu Zeng, Shi Yin1, Dong Wang, "Learning Speech Rate in Speech Recognition", Interspeech 2015.

11. Xi Ma, Xiaoxi Wang, Dong Wang, "LOW-FREQUENCY WORD ENHANCEMENT WITH SIMILAR PAIRS IN SPEECH RECOGNITION", ChinaSIP 2015.

12. Xuewei Zhang, Yiye Lin, Dong Wang, "LASSO-BASED REVERBERATION SUPPRESSION IN AUTOMATIC SPEECH RECOGNITION", ICASSP 2015.

13. Chao Liu, Dong Wang, Zhiyong Zhang, "Pruning Deep Neural Networks by Optimal Brain Damage", Interspeech 2014.

14. Xiangtao Meng, Chao Liu, Zhiyong Zhang, Dong Wang, "NOISY TRAINING FOR DEEP NEURAL NETWORKS", ChinaSIP2014.

通过上述工作，我们在深度学习方法上积累了大累工作经验，包括（1）提出了带噪训练方法和加噪 DAE 方法，可极大提高语音识别系统在噪声环境中的鲁棒性；（2）研究了对低频词的加重方法，提出相似词对方法，在语言模型中对低频词进行加强，自由加入新词；（3）研究了混合识别方法，通过在宿主语言的语言模型中查找相似词对加入外来词，可有效提高混合语言识别性能；（4）研究了语音信号处理和自然语言处理中的转移学习方法，通过转移学习可以有效提高 DNN 模型在小数据集上的性能（5）研究了 DNN 模型优化方法，通过对模型结构进行调整，产生更好的识别效果。

3) 近两年来自然语言处理成果 [SCI 标红]

1. Tang, Guoyu;Xia, Yunqing; Cambria, Erik; Jin, Peng; Zheng, Thomas Fang, "Document representation with statistical word senses in cross-lingual document clustering," International Journal of Pattern Recognition and Artificial Intelligence, v 29, n 2, March 25, 2015,

2. Dong Wang, Thomas Fang Zheng, "Transfer learning for speech and language processing", APSIPA 2015, invited paper

3. Miao Fan, Deli Zhao, Qiang Zhou, Thomas Fang Zheng, Ralph Grishman, "Large Margin Nearest Neighbor Embedding for Knowledge Representation", WI 2015

4. Miao Fan, Qiang Zhou and Thomas Fang Zheng, "Distant Supervision for Entity Linking", PACLIC 2015.

5. Yiqiao Pan, Chao Xing, Dong Wang, "Sphereical Vector for Classification", APSIPA 2015

6. Tianyi Luo, Dong Wang, Rong Liu and Yiqiao Pan, "Stochastic Top-k ListNet", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015 long oral paper) , pp. 676-684, Sep 17-21, 2015. Lisbon, Portugal.

7. Dongxu Zhang, Dong Wang, Rong Liu, "Joint Semantic Relevance Learning with Text Data and Graph Knowledge", ACL 2015, Workshop CVSC.

8. Xi Ma, Xiaoxi Wang, Dong Wang, "Recognize Foreign Low-FrequencyWords with Similar Pairs", Interspeech 2015.

9. Chao Xing, Dong Wang, Chao Liu, Yiye Lin, "Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation", NAACL 2015

10. Miao Fan, Qiang Zhou, Emily Chang, Thomas Fang Zheng, "Transition-based Knowledge Graph Embedding with Relational Mapping Properties", PACLIC 2014.

11. Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, Edward Y. Chang. "Distant Supervision for Relation Extraction with Matrix Completion," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014) (Volume 1: Long

Papers), pp. 839-849, June 23-25, 2014. Baltimore, Maryland. USA [EI:20143718156957]

12. Miao Fan, Qiang Zhou, and Thomas Fang Zheng, “Mining the Personal Interests of Microbloggers via Exploiting Wikipedia Knowledge,” Proceedings of 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014), Kathmandu, Nepal, April 6-12, 2014; A. Gelbukh (Ed.): CICLing 2014, Part II, LNCS 8404, pp. 188-200, 2014, Springer -Verlag Berlin Heidelberg 2014 [WOS:000342990000016,IDS: BB3PT, INSPEC:14269297; EI:20142017719416]

13. Guoyu Tang, Yunqing Xia, Jun Sun, Min Zhang, and Thomas Fang Zheng, “Topic Models Incorporating Statistical Word Senses,” Proceedings of 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014), Kathmandu, Nepal, April 6-12, 2014; A. Gelbukh (Ed.): CICLing 2014, Part I, LNCS 8403, pp. 151-162, 2014, Springer -Verlag Berlin Heidelberg 2014 [WOS:000342989200013,IDS: BB3PO, INSPEC:14269334;EI: 20142017719370]

14. Xia, Yunqing; Tang, Guoyu*; Zhao, Huan Cambria, Erik Zheng, Thomas Fang, "Using word sense as a latent variable in LDA can improve topic modeling", ICAART 2014, Pages: 532-537 [EI: 20142517844063]

上述工作主要集中在对词义和关系进行嵌入式学习(embedding)，即将词和关系投影到连续低维空间，生成词向量或关系向量，进而进行语义计算。我们将研究扩展到文档分类、关系抽取、关系分类、关系补全等各个领域。特别是在工作（9）中，我们研究了不同语言之间词向量的可对应性，证明不同语言可共享语义空间，为语言模型的共享学习提供了依据。

1.2 专利成果

1. 王军;王东;邬晓钧;郑方,基于模型顺序自适应技术的说话人确认系统创建方法, 2015.05, 中国， 201310139096.5
2. 郑方、肖熙、刘林泉、游展、赤羽诚（Makoto Akabane)、陈如新（Ruxin

Chen)、高桥良和（Yoshikazu Takahashi），一种计算机多方言背景的汉语普通话语音识别的建模方法，2011.1，中国，ZL 2008 1 0225354.0

3. 郑方、刘轶、曹文晓、赤羽诚（Makoto Akabane）、陈如新（Ruxin Chen）、高桥良和（Yoshikazu Takahashi），一种汉语语音识别可信度特征值的计算方法，2008.10，中国，ZL 2008 1 0225353.6，
4. 郑方、宋战江；发明专利：一种声纹识别方法；专利号 ZL02149384.7。
5. 郑方、熊振宇、宋战江；发明专利：一种声纹识别方法；专利号 ZL200510059913.1。

1.3 应用成果

1）数据共享

2015 年我们发布了汉语语音数据库 THCHS-30[5]和维吾尔语语音数据库 THUYG-20[6]两个公开数据库，供研究者免费下载使用。我们在 Kaldi 上发布了 THCHS-30 Recipe， 作为和 THCHS-30 数据库配合的基线系统，在数据开放方面做出了尝试。

2）维-汉实时语音翻译系统

2015 年，清华大学、新疆大学与中国民族语文翻译局及捷通华声公司合作，推出维-汉在线语音翻译系统[7,8]，日均访问量过千万次，为维汉交流提供了便利。

3）灵云语音识别平台

2014 年以来，清华大学和捷通华声公司合作推出了灵云语音识别云服务平台，并增加了维吾尔语语音识别服务，日均访问过亿次，经受住了复杂实际场景的检验，积累了构造大型语音识别系统的实际工作经验9。

4）建设银行声密宝

2015 年，清华大学和得意音通公司合作研究了电话银行声纹验证系统，成功用于建设银行。

---

[5] http://data.cslt.org/thchs30/README.html
[6] http://data.cslt.org/thuyg20/README.html
[7] http://www.sinovoice.com/news/company/2015/0210/152.html
[8] http://www.seac.gov.cn/art/2015/6/16/art_34_229794.html
[9] http://www.hcicloud.com/

2. 工作条件（包括已具备的实验条件，尚缺少的实验条件和拟解决的途径，包括利用国家实验室、国家重点实验室和部门重点实验室等研究基地的计划与落实情况）；

项目组由清华大学语音和语言技术中心（CSLT）、新疆大学实验室信号与信息处理重点实验室、西北民族大学民族语言文字信息技术教育部重点实验室联合组成，实验条件具备，研究基地落实。在计算资源方面，CSLT 拥有自己的 100 核计算集群系统，10 个 GPU 计算结点，具有中级运算的能力；新疆大学信号与信息处理重点实验室也配备有28台高配计算机和2台集群服务器； 西北民族大学中国民族语言文字信息技术教育部重点实验室先后投资 1000 万元加强实验室硬件建设，实验室现拥有 IBM 小型机 7 台，高性能 PC 服务器及微机 100 多台，建有专业录音室，拥有专业录音设备及实验语音学研究设备。

2.1 已有的语音库资源

目前清华大学语音语言中心已经积累了相当多的语音和语言数据库资源，包括 863 语音数据库(120 小时)，维吾尔语语音据库(161 小时)，复杂信道语音数据库 (10000 小时以上)。另外，CSLT 是国际中文语言资源联盟的发起单位，目前该联盟已经共享众多数据资源，并准备继续录制一批万人级的跨信道、复杂背景噪音、多说话人的语音数据。根据联盟章程，我们可以使用这些资源用于研究。

新疆大学信号与信息处理重点实验室是新疆大学“211 工程”重点资助的“通信与信息系统”学科的主要基地，现有电话语音语料库(约 100 小时)，新疆人说汉语语料库等都可以用于实验研究。

西北民族大学中国民族语言文字信息技术教育部重点实验室建立了藏语拉萨话朗读式资源库，发音文本来源于人民日报、中央人民广播电台网站的藏语新闻。选择成年男性发音人 45 名，录制了 25292 个句子，共 19.48 小时，选择成年女发音人 36 名，录制 20358 个句子，共 17.62 小时，合计 37.1 小时。建立了藏语拉萨话发音字典，对语料进行了标注。参与了国家高技术研究发展计划(863 计划)项目“以中文为核心的多语言基础资源库的开发和共享”，研发建立了 10 万余词条的藏语词典，各 6000 余字的三大方言发音词典，150 人共 50 小时的藏语三大方言口语语音库，30 万对的汉/藏平等语料库，10G 的文本语料。

2.2 研究队伍

清华大学语音语言中心的前身成立于 1979 年，专业从事语音数字信号处理，现主要从事语音识别、声纹识别、中文自然语言理解等方向，研发成功了许多具有国际先进和国内领先水平的科技成果（有些达到了国际领先水平），获得多项科技成果鉴定、国家科技进步奖、国家发明奖等，并有许多语音产品行销于国内外市场。 2007 年 2 月 CSLT 的前身整合了清华大学信息科学技术学院其他相关课题组，包括电子工程系的成立于 1986 年的语音研发组、电子工程系的语音芯片研发 组、信息技术研究院的未来信息技术（FIT）中心、清华信息科学技术国家实验室的计算机与人工智能研究部等，形成了目前的 CSLT。自成立以来，中心为国 家培养了大量语音和语言领域的优秀人才，现有博士研究生 5 人，硕士研究生 3 人，科研工程师 5 人。

新疆大学信号与信息处理重点实验室的前身是新疆大学智能信息处理重点实验室，建立于 2009 年，该实验室是自治区目前唯一从事信号传输与处理研究的专业实验室。结合 自治区信息化建设的重大需求，面向新疆经济建设与社会发展的需要进行具有民族特色的语音应用技术开发。研发成功了许多针对维吾尔语特点的科研成果，获得了多项 软件登记权等，并有许多语音产品在自治区已产品化。目前，实验室现有博士研究生 5 人，硕士研究生 5 人。

西北民族大学少数民族师生人数众多，少数民族语言语音学研究水平较高，为本项目的开展提供了重要的知识和资源保证。民族语言文字信息技术教育部重点实验室以良好的软硬件环境和扎实的民族语言知识为后盾，以藏语、蒙古语、维吾尔语、汉语、英语等多个语种为研究对象，研究涉及语言学理论、自然语言处理、实验语音学、言语工程等多个领域。西北民族大学民族语言信息处理团队先后承担了国家支撑计划、国家 863 项目、国家自然科学基金重点项目和面上项目。《藏汉双语信息处理系统》获 1999 年国家科技进步二等奖(01-02-005-01)，《藏文视窗平台、字处理软件和藏文网站》2001 年获国家科技进步二等奖（J-220-2-05-R01），多次获得省部级科技进步奖。培养了一批有研发能力，致力于中国民族语言信息技术发展的多学科各民族教学科研骨干，有比较雄厚的技术积累。

3. 正在承担的与本项目相关的科研项目情况（申请人和项目组主要参与者正在承担的与本项目相关的科研项目情况，包括国家自然

科学基金的项目和国家其他科技计划项目，要注明项目的名称和编号、经费来源、起止年月、与本项目的关系及负责的内容等)；

项目名称：　　说话人识别中时变鲁棒的声纹特征研究
NSFC 项目批准号：61271389,
起止日期:　　2013.1-2016.12，
与本项目关系：　　郑方教授为其负责人
状态：　　在研

项目名称：　　语音识别中的稀疏性深度学习
NSFC 项目批准号：61371136,
起止日期:　　2014.1-2017.12，
与本项目关系：　　王东老师为其负责人
状态：　　在研

项目名称：维吾尔单元集优化关键技术研究及其在语音识别中的应用
项目编号：61462085
经费来源：国家自然科学基金委，45 万元
起止年月：2015 年 1 月 1 日 至 2018 年 12 月 31 日
与本项目关系：为本项目提供优化了的维吾尔语语言模型
项目负责人：米吉提·阿不里米提(艾斯卡尔·艾木都拉为主要参与者)
项目状态：在研

项目名称：藏文实体语义关系抽取理论与研究方法
项目编号：61262054
起止日期:　2013-1-1 至 2016-12-31
与本项目关系：于洪志教授为项目负责人
项目状态：在研

项目名称：基于脑电信号的藏语拉萨话韵律认知理论研究

项目编号：61462075

起止日期： 2015-1-1 至 2018-12-31

与本项目关系：于洪志教授为项目负责人

项目状态：在研

4. 完成国家自然科学基金项目情况（对申请人负责的前一个已结题科学基金项目（项目名称及批准号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要（限 500 字）和相关成果的详细目录）。

项目名称：维吾尔文 WEB 舆情挖掘的关键理论技术研究

NSFC 项目批准号：61163033,

起止日期：2012.1-2015.12，

与本项目关系：艾斯卡尔·艾木都拉老师为其负责人

状态：结题

项目名称：少数民族语言文字信息处理共性关键技术研究与示范应用

项目编号：2009BAH41B00

经费来源：国家科技支撑计划项目

起止日期： 2009-8-1 至 2012-12-31

与本项目关系：于洪志为首席专家。

项目状态：结项

项目名称：维吾尔语语素结构及其应用研究

项目编号：61163032

经费来源：国家自然科学基金委，49 万元

起止年月：2012 年 1 月 1 日 至 2015 年 12 月 31 日

与本项目关系：为本项目提供有效的词法分析器

项目负责人：艾克白尔·帕塔尔(艾斯卡尔·艾木都拉为主要参与者)

项目状态：结题

（三）其他需要说明的问题

1. 申请人同年申请不同类型的国家自然科学基金项目情况（列明同年申请的其他项目的项目类型、项目名称信息，并说明与本项目之间的区别与联系）。

项目类型：联合基金项目

项目名称：面向信息安全的维、哈、柯多语言语音识别理论与关键技术研究

与本项目的区别与联系：一是语言数量和所属语系不同，联合基金项目中只考虑了同一个语系（阿尔泰语系突厥语）的维吾尔、哈萨克语、柯尔克孜语；在本项目中考虑了汉藏、阿尔泰语系，同时对资源极度稀缺语言展开研究。二是应用领域不同，联合基金项目以语音检索为主，本项目的研究的应用领域为语音输入法、语音对话等基础语音识别任务。

2. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在同年申请或者参与申请国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，申请或参与申请的其他项目的项目类型、项目名称、单位名称、上述人员在该项目中是申请人还是参与者，并说明单位不一致原因。

3. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在与正在承担的国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，正在承担项目的批准号、项目类型、项目名称、单位名称、起止年月，并说明单位不一致原因。

4. 其他。