资源稀缺蒙语语音识别研究

张爱英 倪崇嘉
(山东财经大学系统科学与信息处理研究所 济南 250014)

摘 要 随着语音识别技术的发展,资源稀缺语言的语音识别系统的研究吸引了更广泛的关注。以蒙语为目标识别语言,研究了在资源稀缺的情况下(如仅有10小时的带标注的语音)如何利用其他多语言信息提高识别系统的性能。借助基于多语言深度神经网络的跨语言迁移学习和基于多语言深度Bottleneck神经网络的抽取特征可以获得更具有区分度的声学模型。通过搜索引擎以及网络爬虫的定向抓取获得大量的网页数据,有助于获得文本数据,以增强语言模型的性能。融合多个不同识别结果以进一步提高识别精度。与基线系统相比,多种系统融合的识别绝对错误率减少12%。

关键词 资源稀缺,多语言深度神经网络,Web语言模型
中图法分类号 TP391.42 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.057

Research on Low-resource Mongolian Speech Recognition

ZHANG Ai-ying NI Chong-jia
(Institute of System Science and Information Processing, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract With the development of speech recognition technology, the research on low-resource speech recognition has gained extensive attention. Taking the Mongolian as the target language, we studied how to use the multilingual information to improve the performance of speech recognition in the low-resource condition, for example, only 10 hours of transcribed speech data are used for acoustic modeling. More discriminative acoustic model can be gotten by using cross-lingual transfer of multilingual deep neural network and multilingual deep bottleneck features. Large amount of web pages can be gotten by using the web search engine and Web crawler, which can help to get large amount of text data for improving the performance of language model. It can further improve the recognition results by fusing different number of recognition results from different recognizers. Comparing the fusion recognition result with the recognition result of baseline system, there are nearly 12% absolute word error rate (WER) reductions.

Keywords Low-resource, Multilingual deep neural network, Web based language model

1 引言

随着计算技术和信息技术的发展,特别是深度机器学习方法的提出及其在语音、图像、视频等领域的成功应用,语音识别系统的性能得到了很大的提高,语音识别技术得到了突飞猛进的发展,同时被应用于各种商业化的识别系统和应用软件中,如Google语音搜索、Bing语音搜索、Siri语音助手、Cortana语音助手、百度语音助手、搜狗语音输入、讯飞语点等。作为人机交互工具,语音正在逐渐改变人与不同设备之间的交互方式。

当前,State-of-the-art语音识别系统需要大量的人工标注语音数据,如Google的英文语音搜索使用了数万小时的语音数据,其语言模型包含数十亿个入口。State-of-the-art语音识别系统需要大量的资源。世界上约有7100多种不同的语言[1],但只有少量的语言可以利用人类语言技术,如当前的Google语音搜索可以提供40多种不同语言的识别系统。世界上96%的人口说的语言只占全部语言的4%左右。在这7100多种不同的语言中,很多语言由于说该种语言的人数较少或缺少书写的文字正在逐渐消失[2],我们将这些语言称为资源稀缺(Low-resource)语言。由于资源稀缺,这些语言利用人类语言技术时面临着很多问题和挑战[3]。资源稀缺语言语音识别已成为语音识别领域中一个非常活跃的研究方向,不仅可以通过研究促进语音识别技术的进一步发展,还可以利用人类语言技术保护这些资源稀缺的语言,有利于文化的保护、传播和传承。因此,资源稀缺语言的语音识别研究具有十分重要的研究价值和意义。

近年来,神经网络展示出改善和提高识别器性能的巨大潜力。用多层感知机(Multilayer Perceptrons, MLP)进行特

到稿日期:2016-09-29 返修日期:2017-01-17 本文受国家自然科学基金(61305027),山东省自然科学基金(ZR2011FQ024),山东省高等学校科技计划(J17KB160)资助。

张爱英 女,讲师,主要研究方向为模式识别、数字信号处理等,E-mail:ayzhang@sdufe.edu.cn;**倪崇嘉** 男,博士,主要研究方向为模式识别、语音识别、音频分类等。

征抽取，如 Tandem 特征[4-5] 和 Bottleneck 特征[6]，研究结果表明：MLP 特征具有较好的区分能力，对说话人及环境的变化有较强的鲁棒性，并具有一定的语言独立性。对资源稀缺语言而言，这些特性可以使开发者利用少量资源稀缺的目标语言数据和大量资源丰富的源语言数据来搭建更加有效的识别系统。Thomas 等和 Vesely 等[7-8] 研究了如何利用多语言数据来抽取资源稀缺语言的特征。Vu 等基于国际音标（International Phonetic Alphabet，IPA）音位映射，利用多语言的多层感知机来初始化资源稀缺语言的多层感知机以提高和改善资源稀缺语言识别系统的性能[9]。在构建这些深度神经网络时，其训练的目标不尽相同，可以利用 Senones 或通用的音子集等。Miao 等则利用多语言的 Maxout 深度神经网络和卷积神经网络来抽取特征[10]。

利用深度神经网络进行多语言和跨语言迁移学习的研究越来越受到人们的重视[11-13]，特别是对资源稀缺语音识别系统的研究。文献[11，14-15]提出了多种不同的深度神经网络框架。Huang 等[15] 描述了共享隐层的多语言深度神经网络训练（SHL-MDNN）框架。Ghoshal 等提出了另外一种多语言深度神经网络的训练方法[14]；在该方法中，首先，所有的隐层利用无监督的深度信念网络（Deep Belief Network，DBN）进行预训练，然后选中一种语言，添加输出层及 Softmax 层，其中输出层所含的单元个数与该种语言的 Senones 的数目一致，且被随机初始化，利用该种语言数据对该深度神经网络进行细调（Fine-tuning），之后，该深度神经网络的输出层和 Softmax 层被移除；另一种新的语言继续该过程，直到所有的语言都进行了上述的操作。而 SHL-MDNN 框架则是多任务学习的一种特殊形式。多任务学习使 DNN 训练时更有效，其原因在于：1）它提供了支持所有任务局部优化的表示偏移；2）它能够缓解由于不同语言共享隐层而引起的过拟合问题，使得共享隐层的参数可以得到很好的训练；3）它支持并行的方式；4）它可以提高系统的泛化能力。多任务学习能够提高所涉及任务的泛化能力，在语音识别领域已用于更泛化的深度神经网络的多目标训练。Lu 等[16] 利用多任务学习来改善带噪数字串识别的鲁棒性，并利用只有一个隐层的递归神经网络（Recurrent Neural Network，RNN）来识别数字串。不同于现有的方法，其训练的神经网络可以同时分类数字、增强语音和识别说话人的性别；在 Aurora 数据集上的实验表明，该方法能够减少 50％的相对错误率。Seltzer 等在 TIMIT 数据集上利用多任务学习来改善 DNN-HMM 系统的音子识别性能；训练 DNN 时，除利用单音子的 3 个状态（共 183 个）作为训练的目标，同时还利用音子、状态上下文及音子上下文作为训练的目标。其实验结果表明：通过添加音子作为深度神经的训练目标，其训练获得的 DNN 不会影响音子的识别结果；而通过添加音子上下文作为训练目标，其训练获得的 DNN 可以提高音子的识别正确率，减少错误率。实验证明，多任务的 DNN 训练通过在不同任务中学习共同的结构可以提高模型的泛化能力[17]。Chen 等通过联合训练 DNN 提高系统的泛化能力，利用 Tri-phone 和 Tri-grapheme 作为训练目标，在 3 种资源稀缺的南非语言（南非荷兰语、赛索托语、斯瓦特语）上的实验结果表明，基于多任务学习获得的 DNN 比利用单任务学习获得的 DNN 模型的错误率减少了 5％～13％[18]。

针对资源稀缺的蒙语语音识别，本文提出利用多语言和跨语音信息来提高蒙语语音识别系统的性能。将 17 种不同语言的语音数据用于训练不同的 Hybrid 多语言深度神经网络和多语言 Bottleneck 深度神经网络，以期通过这些跨语言信息的应用来提高蒙语的声学建模能力。同时针对资源稀缺语言的文本数据的不足，提出了利用搜索引擎和定向抓取网站的方法来获取大量的文本数据，对获取的数据进行清洗后用于训练语言模型。实验结果表明，通过 Web 获取文本数据的方法不仅可以扩充词表，而且可以提高语言模型的性能。最后，通过融合不同识别系统的识别结果可以进一步改善系统的识别结果。对 10 小时的蒙语语音识别的实验表明，结合了这些技术的识别系统的性能有了大幅度的提高，绝对错误率降低约 12％。

本文第 2 节介绍了多语言的深度神经网络和跨语言的迁移学习；第 3 节介绍了基于 Web 的数据获取和 Web 语言模型；第 4 节给出了实验设置及结果；第 5 节给出了实验结果的进一步分析；最后总结全文并展望未来。

2　多语言深度神经网络

2.1　Hybrid 多语言深度神经网络及迁移学习

共享隐层的多语言深度神经网络训练（SHL-MDNN）框架[15] 描述了一种有效的多语言深度神经网络训练方法。在这种框架中，隐层被所有语言共享，而 Softmax 层不被共享，即不同的语言有不同的 Softmax 层。该多语言深度神经网络的核心思想就是：深度神经网络的隐层被看成级联的特征抽取器，且仅有输出层与分类相关。从多语言深度神经网络中抽出的共享隐层可以看作是由多种不同语言训练的智能特征抽取模块，它们携带着丰富的信息，可以用于辨别新语言中的音子信息。通过将这些共享的隐层迁移到新的语言中，不仅可大大减少训练数据量，还可使跨语言模型迁移的过程变得简单，且不同的语言之间可以有效地迁移。图 1 给出了该种共享隐层的多语言深度神经网络的结构图。

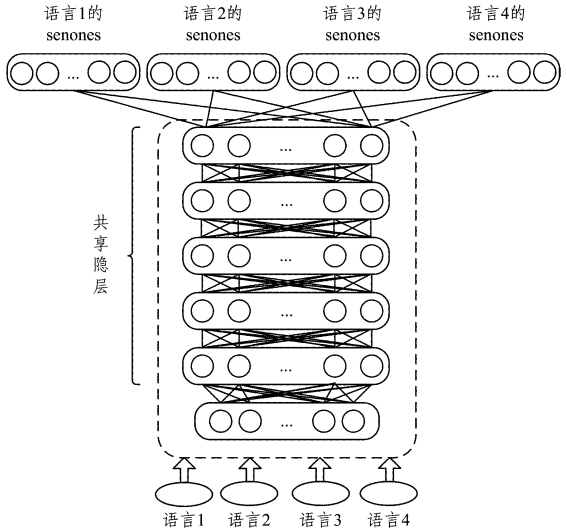

图 1　共享隐层的多语言深度神经网络的结构图

基于 SHL-MDNN 的深度神经网络既可以用于训练 Hy-

brid 多语言深度神经网络，也可以用于训练含有 Bottleneck 层的多语言 Bottleneck 特征抽取器(Multilingual Bottleneck Extractor)。之前的研究也表明：基于多语言 Bottleneck 特征的跨语言迁移比多语言 Hybrid 深度神经网络的迁移更有效[19]。

对于 Hybrid 的多语言深度神经网络的跨语言的迁移学习，原来的语言相关的 Softmax 层被移除，新的与目标语言相关的 Softmax 层被置于共享隐层之上，与共享隐层进行连接。目标语言相关的 Softmax 层的 Senones 的数目与目标语言数据中用于切分该目标数据 Hybrid 的 GMM-HMM 系统或 DNN-HMM 系统的 Senones 的数目一致。利用目标语言数据，该深度神经网络被细调(Fine-Tuning)。细调的神经网络层数不定，可以根据实际识别系统的识别结果选择一个最好的深度神经网络。

2.2 多语言 Bottleneck 深度神经网络

多语言 Bottleneck 深度神经网络可以按照 Hybrid 多语言深度神经网络的跨语言学习那样进行自适应。然而，当 Bottleneck 层处于较深的层次(从上到下)以及只有较少的目标语言数据(如 10 小时的目标语言数据)时，该自适应方法基本无效。

利用该多语言 Bottleneck 特征提取器(Multilingual Bottleneck Feature Extractor)抽取 Bottleneck 特征，并将该特征与用于抽取该 Bottleneck 特征的原始特征相拼接，组成新的特征以训练 Hybrid 的深度神经网络。用该方法构造的 Hybrid 深度神经网络比仅对 Hybrid 多语言深度神经网络进行迁移学习获得的深度神经网络更有效[19]。

3 Web 文本数据获取与 Web 语言模型

与获取带标注的目标语言语音数据相比，通过互联网可以很方便地获得目标语言相关的文本数据。首先，不同的搜索引擎可以十分方便地用于搜索包含特定文本的页面。其次，可以利用网络爬虫实现对某些目标语言站点的定向数据抓取。通过这两种方法可以获得大量的包含目标语言的页面[20-23]。

利用搜索引擎或网络爬虫的定向抓取获得这些页面之后，可以从这些页面中抽取文本数据。由于抽取的页面文本数据中包含了很多噪音数据，因此在利用这些数据训练 Web 语言模型之前，需要对其进行预处理：

首先，将蒙语进行转换，即将传统的蒙文转为拉丁蒙文。

其次，根据蒙文的标点符号，将文本数据分成一个个句子，即一个句子占据一行。

最后，将数字、日期等阿拉伯数字转为拉丁蒙文，并清除一些非拉丁蒙文的字符。

基于这些数据来训练 Web 语言模型。在训练语言模型时，仅让高频词出现在该语言模型中。由于该 Web 语言模型与其应用的领域可能不一致，该 Web 语言模型与利用训练脚本获得的语言模型进行插值，以获得更好的语言模型。对插值的权值根据评测数据的文本进行调优以获得更好的识别效果。

4 实验

4.1 实验设置

将多个带标注的不同语言的语音数据用于训练多语言的深度神经网络，表 1 列出了这些语言数据。其中，源语言数据用于训练 Hybrid 多语言的深度神经网络或多语言 Bottleneck 深度神经网络。目标语言数据是带标注的 10 小时蒙语数据。所有语言的语音数据都是自然口语风格的对话语音数据，并以 8000Hz 的频率进行采样，以 16 位的长度进行量化编码。对话的内容涉及日常生活，如聊天、购物、看医生、吃饭等。

表 1 中，孟加拉语、克里奥耳语、老挝语、库尔德语、祖鲁语、哈萨克语、立陶宛语、巴拉圭语、阿姆哈拉语、爪哇语等 10 种语言用于训练一个多语言深度神经网络。广东话、阿萨姆语、普什图语、土耳其语、塔加拉族语、祖鲁语、立陶宛语、巴拉圭语、阿姆哈拉语等 9 种语言用于训练一个多语言深度神经网络。克里奥耳语、伊博语、阿姆哈拉语、卢欧语、广东话、普什图语、土耳其语、塔加拉族语等 8 种语言用于训练一个多语言的 Bottleneck 特征提取器。

表 1 不同语言语音数据

类型	语言	大小/h	类型	语言	大小/h
源语言	孟加拉语 (Bengali)	87	源语言	爪哇语 (Javanese)	59
	克里奥耳语 (Creole)	83		伊博语 (Igbo)	56
	老挝语 (Lao)	86		卢欧语 (Dholuo)	54
	库尔德语 (Kurdish)	51		广东话 (Cantonese)	175
	祖鲁语 (Zulu)	84		土耳其语 (Turkish)	107
	哈萨克语 (Kazakh)	54		塔加拉族语 (Tagalog)	115
	立陶宛语 (Lithuanian)	54		普什图语 (Pushto)	111
	巴拉圭语 (Paraguayan)	55		阿萨姆语 (Assamese)	74
	阿姆哈拉语 (Amharic)	56	目标语言	蒙古语 (Mongolian)	10

在选择这些语言进行分组时，考虑了这些语言的地域分布及语言特点，如不同的国家、不同洲、不同语系等。

在训练这些多语言深度神经网络时，每种语言的 Senones 的数目大约是 3500～4500，共有 6 个隐层，除 Bottleneck 层含有 42 个隐层单元之外，其他每一个隐层所含的隐层单元的个数是 2048。每种语言的 GMM-HMM 系统用于强制对齐数据，产生强制切分以用于训练多元深度神经网络。Kaldi 工具用于训练多语言的深度神经网络以及不同语言的 GMM-HMM 系统[24]。13 维的 MFCC 特征以及它们的一阶、二阶差分(共 39 维)用于训练每种语言的 GMM-HMM 系统。40 维的 FBank 特征以及 3 维 Kaldi 的 Pitch 特征(共 43 维)用于训练多语言深度神经网络。在抽取这些特征时，设定 25ms 的 Hamming 窗，且窗移为 10ms。交叉熵(Cross－entropy)准则用于训练多语言深度神经网络。单一目标语言的深度神经网络或者跨语言迁移的神经网络都是先基于交叉熵(Cross－entropy)准则训练，然后再按照 sMBR 准则进行序列训练

(Sequence Training)。训练多语言深度神经网络的初始学习速率设置为 0.08，并按指数准则减少。

实验中，利用 Bing 和 Google 来搜索包含特定词的页面，并在 2016 年 4 月到 2016 年 5 月期间利用 Scrapy 对人民网(蒙古文版)、中国蒙古语信息网(蒙古文版)、好乐宝博客网等站点进行定向抓取，总共获取了 20 万个页面，涉及的内容比较广泛，包括新闻、博客、小说、广告等，语言风格与自然口语有较大的不同。数据清洗完成后，有 877M 文本数据。这些数据将用于训练 Web 语言模型。将 10h 的自然口语风格的带标注的蒙语作为目标语言，用于训练基线系统以及进行跨语言的迁移学习。10h 蒙语训练的 GMM-HMM 系统用于强制切分，其包含 2009 个 Senones。另外 10h 的自然口语风格的蒙语数据用于系统评测。10h 的训练数据的脚本包含有 8507个词，基于训练脚本训练的 3-gram 语言模型的评测脚本的困惑度是 154.3。利用 Web 数据训练的 3-gram 语言模型与训练数据脚本的语言模型进行插值，且该插值的权值根据评测脚本来优化以获得新的 Web 语言模型。该语言模型包含 23989 个词，其评测脚本的困惑度是 127.7。利用词错误率(Word Error Rate，WER)来评测系统的性能。

4.2　实验结果

表 2 列出了基于目标语言而构建的基线系统的性能。从表 2 中可以看到：1)由于仅有 10h 的训练数据通过其训练获得的系统的性能不是很好。在基于训练脚本的语言模型进行解码的情况下，WER 是 73.1%。2)利用 Web 数据可以提高识别系统的性能。在基于 Web 语言模型进行解码的情况下，WER 是 71.8%。与基于训练脚本语言模型进行解码的结果相比，系统的 WER 有 1.3%绝对错误率减少。

表 2　基线系统的识别结果

系统	WER/%	
	训练脚本语言模型	Web 语言模型
基线系统	73.1	71.8

在进行跨语言的迁移学习时，为了获得最佳的识别效果，对不同个数的隐层进行了细调。利用 10 种语言训练获得的多语言深度神经进行迁移学习，通过细调不同层数的深度神经网络获得的识别系统的性能如图 2 所示。在图 2 中，深度神经网络仅采用交叉熵进行训练，且仅基于训练脚本语言模型进行解码。细调时初始学习速率设置为 0.02。

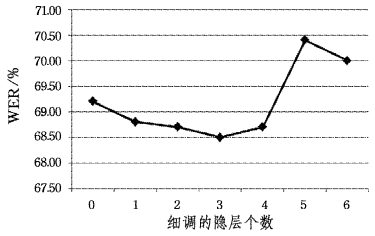

图 2　细调不同隐层个数的神经网络在评测集上获得的识别系统的性能

从图 2 可以看到，系统在细调 3 个隐层和输出层时获得最佳的识别性能。同样地，在利用 9 种语言获得的多语言深度神经网络进行迁移学习时，也对 3 个隐层进行细调。

表 3 列出了基于多语言深度神经网络进行跨语言迁移学习获得的深度神经网络系统的性能。表 3 中，所有深度神经网络都经过 sMBR 准则进行序列训练。

表 3　基于多语言深度神经网络迁移的识别系统的识别结果

编号	系统描述	WER/%	
		训练脚本语言模型	Web 语言模型
系统 1	基于 10 种语言的深度神经网络迁移的神经网络	65.1	62.5
系统 2	基于 9 种语言的深度神经网络迁移的神经网络	66.1	63.8
系统 3	结合基于 8 种语言的深度 Bottleneck 抽取进行特征训练获得的深度神经网络	63.3	61.1

由表 3 可知，无论是基于 Hybrid 的多语言深度神经网络进行迁移，还是利用多语言的 Bottleneck 特征抽取器提取特征，都可以提高识别系统的性能。在利用 Web 语言模型进行解码的情况下，对比基线系统，绝对错误率降低了 8%～10.7%。对比基于 Hybrid 多语言深度神经网络的迁移获得的神经网络与结合多语言的 Bottleneck 特征抽取器提取的特征进行训练获得的深度神经网络可知，后者能获得更好的识别效果。

由于在构建表 3 中的不同的识别系统时，其声学模型有很大的不同，使得不同识别系统的识别结果有较大的互补性，因此融合不同系统的识别结果可进一步提高系统的性能。表 4 列出了利用 ROVER 对不同识别系统的识别结果进行融合后获得的结果[24]。

表 4　不同识别系统识别结果的融合

融合	WER/%	
	训练脚本语言模型	Web 语言模型
系统 1 ＋ 系统 2	64.1	61.2
系统 1 ＋ 系统 3	62.2	60.4
系统 2 ＋ 系统 3	63.0	60.5
系统 1 ＋ 系统 2 ＋ 系统 3	61.6	59.8

从表 4 可以看到，无论是两个系统的融合，还是 3 个系统的融合，都可以进一步提高系统的识别效果。与基线系统相比，在 Web 语言模型解码的情况下，融合 3 个识别系统的识别结果后 WER 降低了约 12%。

5　实验结果的进一步分析

研究表明[26]，如果用于训练多语言深度神经网络的语言与目标语言越相似，基于这些语言训练获得的多语言深度神经网络在利用跨语言的迁移学习获得深度神经网络之后，越能够提高目标语言的识别性能。基于语种识别(Language Identification，LID)的方法测试了所选的不同语言与目标语言(蒙语)的相似程度[26]。图 3 给出了该实验的结果。

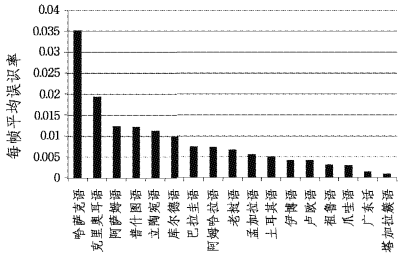

图 3　蒙语的每帧被误识为其他语言的统计结果

从该实验可以看到，在所选择的这些语言中，蒙语的每帧被误识为哈萨克语的概率最高，表明该语言与蒙语有较高的相似性。由语言学可知二者都属于阿尔泰语系，且哈萨克语的形成深受蒙语等语言的影响[27]。

结束语 由于缺少足够多的带标注的语音数据，构建 State-of-the-art 的蒙语语音识别系统是很困难的。多语言和跨语言信息的利用有助于提高识别系统的声学建模能力，从而进一步提高识别系统的性能。基于多语言神经网络的迁移学习以及基于多语言深度 Bottleneck 特征提取器抽取的特征都可以用于提高资源稀缺蒙语的语音识别系统的性能。而基于互联网抓取到的大量页面可以进一步改善和提高系统的性能。不同识别系统的融合可以在上述基础上进一步提高识别系统的性能。从实验中还看到，由于带标注的目标语言语音数据不足，使得系统的识别性能还不能够达到 State-of-the-art 的效果。在以后的工作中，我们将利用增加数据或半监督机器学习的方法来弥补该不足，期望这些方法可以提升系统的性能。另一方面，基于多语言语音数据训练多语言深度神经网络时，仅考虑了不同的语言组以及从不同地域中选择不同的语言来训练多语言深度神经网络，未考虑目标语言的特性，也未考虑这些语言对目标语言的影响。在今后的工作中，将结合目标语言的特点，选择对目标语言最有效、最能提高目标语言识别性能的语言来训练或快速自适应多语言深度神经网络，并分析这些不同的语言数据对目标语言识别性能的影响。

参 考 文 献

[1] Ethnologue [OL]. http://www.ethnologue.com.

[2] Summer Institute for Linguistics Ethnologue Survey 1999[OL]. https://afrobranding.wordpress.com/tag/summer-institute-for-linguistics-sil-ethnologue-survey.

[3] BESACIER L,BARNARD E,KARPOV A,et al. Automatic Speech Recgnition for Under-resourced Languages: A Survey [J]. Speech Communication,2014,56(1):85-100.

[4] HERMANSKY H,SHARMA S. Temporal Patterns (TRAPS) in ASR of Noisy Speech [C]// Proc. of ICASSP 1999. 1999:289-292.

[5] HERMANSKY H,SHARMA S,JAIN P. Data-derived Nonlinear Mapping for Feature Extraction in HMM[C]// Proc. of ASRU. 1999.

[6] GRÉZL F,KARAFIA M,KONTAR S,et al. Probabilistic and Bottle-Neck Features for LVCSR of Meetings[C]// Proc. of ICASSP 2007. 2007:757-760.

[7] THOMAS S,GANAPATHY S,HERMANSKY H. Multilingual MLP features for Low resource LVCSR Systems[C]// Proc. of ICASSP 2012. 2012:4269-4272.

[8] VESELY K,KARAFIAT M,GREZL F,et al. The Language-independent Bottleneck Features[C]// Proc. of SLT 2012. 2012:336-341.

[9] VU N T,BREITER W,METZE F,et al. An Investigation on Initialization Schemes for Multilayer Perceptron Training Using Multilingual Data and Their Effect on ASR Performance[J]. Interspeech,2012,26(5):25681-25689.

[10] MIAO Y,METZE F. Improving Language-Universal Feature Extraction with Deep Maxout and Convolutional Neural Networks [C]// Proc. of Interspeech 2014. 2014:800-804.

[11] YU D,DENG L. Automatic Speech Recognition-A Deep Learning Approach[M]. Springer Press,2014.

[12] DAHL G E,YU D,DENG L,et al. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech RecognitionJ]. IEEE Transactions on Audio,Speech,and Language Processing,2012,20(1):33-42.

[13] HINTON G,DENG L,YU D,et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition [J]. IEEE Signal Processing Magazine,2012,29(6):82-97.

[14] GHOSHAL A,SWIETOJANSKI P,RENTALS S. Multilingual Training of Deep Neural Networks[C]// Proc. of ICASSP 2013. 2013:7319-7323.

[15] HUANG J T,LI J,YU D,et al. Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers [C]// Proc. of ICASSP 2013. 2013:7304-7308.

[16] LU Y,LU F,SEHGAL S,et al. Multitask Learning in Connectionist Speech Recognition[C]// Proc. of Australian International Conference on Speech Science and Technology. 2004.

[17] SELTZER M L,DROPPO J. Multi-task Learning in Deep Neural Networks for Improved Phoneme Recognition[C]// Proc. of ICASSP 2013. 2013:6965-6969.

[18] CHEN D,MAK B,LEUNG C C,et al. Joint Acoustic Modeling of Triphones and Trigraphemes by Multi-task Learning Deep Neural Networks for Low-resource Speech Recognition [C]// Proc. of ICASSP 2014. 2014:5592-5596.

[19] XU H,DO V H,XIAO X,et al. A Comparative Study of BNF and DNN Multilingual Training on Cross-lingual Low-resource Speech Recognition [C]// Proc. of Interspeech 2015. 2015:2132-2136.

[20] MENDELS G,COOPER E,SOTO V,et al. Improving Speech Recognition and Keyword Search for Low-resource Languages Using Web Data[C]// Proc. of Interspeech 2015. 2015:829-833.

[21] CUCU H,BUZO A,BESACIER L,et al. SMT-based ASR Domain Adaptation Methods for Under-resourced Languages: Application to Romanian [J]. Speech Communication,2014,56(1):195-212.

[22] OFLAZER K,EL-KAHLOUT I D. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation [C] // Proc. of Statistical Machine Translation Workshop at ACL 2007. 2007:25-32.

[23] XIE C,GUO W,HU G,et al. Web Data Selection Based on Word Embedding for Low-resource Speech Recognition [C]// Proc. of Interspeech 2016. 2016:1340-1344.

[24] POVEY D,GHOSHAL A,BOULIANNE G,et al. The Kaldi Speech Recognition Toolkit [C]// Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. 2011.

[25] STOLCKE A. SRILM-An Extensible Language Modeling Toolkit [C]// Proc. of ICSLP 2002. 2002.

[26] ZHANG Y,CHUANGSUWANICH E,GLASS J. Language ID-based Training of Multilingual Stacked Bottleneck Features [C]// Proc. of Interspeech 2014. 2014:1-5.

[27] 曹道巴特尔. 喀喇沁蒙古语研究[M]. 北京：民族出版社，2007.