基于多语言语音数据选择的资源稀缺蒙语语音识别研究

张爱英

(山东财经大学数学与数量经济学院 济南 250014)

摘 要 利用多语言信息可以提高资源稀缺语言识别系统的性能。但是,在利用多语言信息提高资源稀缺目标语言识别系统的性能时,并不是所有语言的语音数据对资源稀缺目标语言语音识别系统的性能提高都有帮助。文中提出利用长短时记忆递归神经网络语言辨识方法选择多语言数据以提高资源稀缺目标语言识别系统的性能;选出更加有效的多语言数据用于训练多语言深度神经网络和深度Bottleneck神经网络。通过跨语言迁移学习获得的深度神经网络和通过深度Bottleneck神经网络获得的Bottleneck特征都对提高资源稀缺目标语言语音识别系统的性能有很大的帮助。与基线系统相比,在插值的Web语言模型解码条件下,所提系统的错误率分别有10.5%和11.4%的绝对减少。

关键词 数据选择,资源稀缺,多语言深度神经网络,深度Bottleneck神经网络

中图法分类号 TP391.42 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.09.052

Research on Low-resource Mongolian Speech Recognition Based on Multilingual Speech Data Selection

ZHANG Ai-ying

(School of Mathematic and Quantitative Economics, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract The performance of low-resource speech recognition system is improved by the multilingual information. However, when the multilingual information is used to improve the performance of low-resource automatic speech recognition system, not all of the multilingual speech data could be utilized to improve the performance of low-resource automatic speech recognition system. In this paper, a data selection method which is based on long short-term memory recurrent neural network based language identification was proposed and used to improve the performance of low-resource automatic speech recognition system. More efficient multilingual speech data are selected and used to train multilingual deep neural network and deep Bottleneck neural network. The deep neural network model obtained by using transfer learning and the Bottleneck features extracted by using the deep bottleneck neural network are both helpful to improve the performance of low-resource target language speech recognition system. Comparing with the baseline system, there are 10.5% and 11.4% absolute word error rate reductions under the condition of interpolated web based language model for decoding.

Keywords Data selection, Low-resource, Multilingual deep neural network, Deep Bottleneck neural network

1 引言

随着深度学习的发展、计算机计算性能的提高和大规模训练数据的广泛应用,语音识别系统的性能得到了极大的提高,语音识别技术得到了飞速发展。移动端的语音搜索、语音输入、语音助手以及智能家居中的智能音箱等与语音相关的产品和服务层出不穷。语音,作为一种人机交互中更人性化的工具,正改变着人与不同设备之间的交互方式。

构建最先进的(state-of-the-art)语音识别系统需要大量的人工标注语音数据。一方面,标注这些语音数据需要大量的人力和物力;另一方面,存在大量的原始语音数据时,才可以进行标注。对于只有少数人使用的语言,这些资源(包括数据资源和人力资源)都是稀缺的。当前世界上约有7100多种活跃的语言[1-2],但只有极少数的语言能够利用人类语言技术进行处理。因此,利用人类语言技术对资源稀缺语言进行技术处理(如进行语音识别)的研究引起了越来越多学者的关注。

利用深度神经网络进行多语言和跨语言迁移学习来提高资源稀缺语言语音识别系统的性能是一种行之有效的方法,也受到了研究人员的广泛关注。目前已出现了多种不同的多

到稿日期:2017-11-17 返修日期:2018-01-05 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项目(J17KB160)资助。

张爱英(1980—),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(通信作者)。

语言和跨语言迁移学习的深度神经网络框架[3-8]。一种多语言和跨语言迁移学习框架是 Huang 等提出的共享隐层多语言深度神经网络(Shared Hidden Layer-Multilingual Deep Neural Network,SHL-MDNN)训练框架及跨语言迁移学习方法[6]。该方法中,所有的隐层是不同语言之间共享的,而输出层则是语言相关的,即每种语言都有与该语言相关的输出层。另外一种多语言和跨语言的迁移学习框架是由 Ghoshal 等提出的一种逐步训练的多语言深度神经网络的训练方法[9]。该方法首先用基于无监督深度信念网络(Deep Belief Network,DBN)预训练的网络来初始化一个深度神经网络模型,然后在其输出层被随机初始化后随机选择一种语言对整个神经网络进行有监督的训练,训练完成之后将与该语言相关的输出层和 Softmax 层移除。当所有的语言都进行了上述处理之后,训练结束。

Lu 等利用多任务学习的方式使带噪数字串识别系统的鲁棒性得到了大幅提高[10]。将三音子(Tri-phone)和三音素(Tri-grapheme)同时作为深度神经网络的训练目标,通过多任务的联合训练可以提升识别系统的泛化能力,从而增强系统的识别效果[11]。基于 Babel 项目的多语言数据,Cui 等更加细致地探讨了多语言数据对资源稀缺语言的语音识别系统和关键词检索系统性能的影响[8]。已有工作[12]研究了利用多语言信息进行跨语言迁移学习,与本文工作的不同点在于:已有工作仅利用多语言信息来提高目标语言的识别性能,而本文工作对如何选择以及选择何种多语言进行跨语言的迁移学习进行了研究。

在利用多语言深度神经网络进行跨语言的迁移学习来提升资源稀缺语言语音识别系统的性能时,并不是所有语言的语音数据对资源稀缺语言的语音识别系统性能的提高都有促进作用[3]。因此,选择对目标语言语音识别性能的提高最有益的多语言语音数据来训练多语言深度神经网络进行跨语言的迁移学习显得尤为重要[3,13-14]。文中用不同的方法选择多语言语音数据来训练多语言深度神经网络,以使其能够更有效地迁移到资源稀缺的目标语言上;针对资源稀缺的蒙语语音识别的问题,提出了基于语言辨识的方法来选择多语言语音数据;同时,为了提高不同语言辨识系统的效果,长短时记忆递归深度神经网络被用于训练语言辨识模型。基于资源稀缺的蒙语语音识别的实验结果表明,基于语言辨识的方法选择出的多语言语音数据能够训练获得更有效的多语言深度神经网络,从而更有效地进行迁移学习。与基于其他语言训练的多语言深度神经网络迁移获得的识别系统相比,在插值的 Web 语言模型解码的条件下,所提系统的错误率有 1.2%的绝对减少。与基线系统相比,在插值的 Web 语言模型解码的条件下,所提系统的错误率有 10.5%的绝对减少。

本文第 2 节对基于长短时记忆递归神经网络语言辨识方法的多语言语音数据的选择进行详细的介绍;第 3 节对共享隐层的多语言深度神经网络以及跨语言迁移学习进行了介绍;第 4 节给出实验设置、实验结果的描述与分析;最后总结全文,并对下一步工作进行展望。

2　基于长短时记忆递归神经网络语言辨识方法的多语言语音数据的选择

在利用多语言语音数据训练多语言深度神经网络进行跨语言的迁移学习时,一种直观的想法是:利用与该目标语言相近的数据来训练多语言深度神经网络。这样,在基于该多语言深度神经网络进行跨语言的迁移学习时,由于语言的相似性,在迁移学习之后获得深度神经网络对资源稀缺语言会更有效。采用语言辨识的方法可以衡量这些多语言源语音数据与资源稀缺的目标语言语音数据的相似性。

语言辨识也即识别语言的类别。许多不同的机器学习方法可以构建语言辨识模型。长短时记忆递归深度神经网络因能刻画更长时间的依赖,具有记忆性,在语言辨识时具有最好的(state-of-the-art)效果[15]。

长短时记忆递归神经网络(Long Short-term Memory Recurrent Neural Network,LSTM-RNN)计算从输入的特征序列 $x=(x_1,x_2,\cdots,x_T)$ 到输出的序列 $y=(y_1,y_2,\cdots,y_T)$ 之间的映射。对于时间帧,$t=1,\cdots,T$,神经网络的激活可以利用式(1)－式(6)计算获得。

$$i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i) \tag{1}$$

$$f_t=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f) \tag{2}$$

$$c_t=f_t\cdot c_{t-1}+i_t\cdot tanh(W_{xc}x_t+W_{hc}h_{t-1}+b_c) \tag{3}$$

$$o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}c_t+b_o) \tag{4}$$

$$h_t=o_t\cdot tanh(c_t) \tag{5}$$

$$y_t=\phi(W_{yh}h_t+b_y) \tag{6}$$

其中,W 表示连接不同门的权值矩阵;b 是偏值矢量;W_{ci},W_{cf},W_{co} 是 peephole 连接的对角型矩阵;σ 是 sigmoid 函数;i,f,o,c 分别表示输入门、遗忘门、输出门和记忆细胞激活矢量,它们都有与记忆细胞输出相同的维数;"$\cdot$"表示矢量的逐元素相乘,$tanh$ 是 tanh 激活函数,ϕ 表示 Softmax 输出激活函数。

当一层神经网络的输出再作为输入而输入到神经网络时,就构成了深度神经网络。

设 N 是语言的数目,因为在训练 LSTM-RNN 模型时,所有语言语音数据的静音部分的标签都是同一个,比如记为 0,而该语句语言数据帧部分的标签则标注为该语言的标签,如第 i 种语言,其类别标签标注为 i,所以在利用 LSTM-RNN 构建语言辨识模型时,该 LSTM-RNN 的输出所对应的标签数目是 $N+1$。

设 $x=(x_1,x_2,\cdots,x_T)$ 是输入的多语言数据句子的语音特征,S 是所有多语言语音句子的集合,$p(L\mid x_t)=(p(l_0\mid x_t),p(l_1\mid x_t),\cdots,p(l_N\mid x_t))$ 是输入的特征 x_t 在第 t 帧的后验概率矢量,其中,$p(l_i\mid x_t)$ 是输入特征 x_t 为语言 l_i 的后验概率。则定义如下函数用于数据选择:

$$f(S)=\sum_{s\in S}\log(Proj_{i_k}(\frac{1}{T(s)}\sum_{t=1}^{T}p(L\mid x_t^s))) \tag{7}$$

其中,$T(s)$ 表示输入句子 s 所包含的帧的数目;x_t^s 表示输入句子 s 的第 t 帧输入特征;$Proj_{i_k}(\cdot)$ 为投影函数,该函数可以得到矢量的第 i_k 个分量,i_k 是目标语言。为了减少帧层次的误分,将整个句子层次后验概率的平均用于句子选择。当

利用式(7)选择句子时,具有较高误分率的句子将被选择。

3 多语言深度神经网络

3.1 多语言深度神经网络及跨语言迁移学习

共享隐层多语言深度神经网络(Shared Hidden Layer-Multilingual Deep Neural Network,SHL-MDNN)是一种多语言的深度神经网络[6]。该种类型的深度神经网络中所有的隐层是共享的,而输出层是语言相关的,即不同语言对应着不同的输出层。从模式识别的角度看待 SHL-MDNN,可以把共享隐层看作一个级联的特征抽取器,而将与语言相关的输出层和 Softmax 层看作分类器,其对经过层层抽象获得的高层次特征进行分类。

在训练共享隐层多语言深度神经网络时,不同语言之间共享这些隐层,从而使得共享隐层编码了丰富的可用于辨识不同语言的音子信息,也使得输入特征在经过共享隐层的逐层抽象之后对不同的语言更具区分能力。

在利用共享隐层多语言深度神经网络进行跨语言迁移学习时,首先,SHL-MDNN 与语言相关的输出层和相应的 Softmax 层被移除;然后,初始化一个与目标语言相关的且只有一层的神经网络。该神经网络的输出层所含节点数目与目标语言所对应的输出一致,输入层所含节点数目与共享隐层的输出一致。例如,如果目标语言的输出是 GMM-HMM 系统所对应的 senones,则该神经网络的输出层所含的节点数目就是其 GMM-HMM 系统所含 senones 的数目。共享隐层与神经网络连接,构建一个深度神经网络。最后,利用目标语言数据和较小的学习速率对该深度神经网络进行细调(Fine-Tuning)。在细调该深度神经网络时,细调的深度神经网络的层数以及学习速率的大小需要根据目标语言数据的数量和在目标语言上的识别效果确定。

图 1 给出了该种共享隐层多语言深度神经网络 SHL-MDNN 的结构。

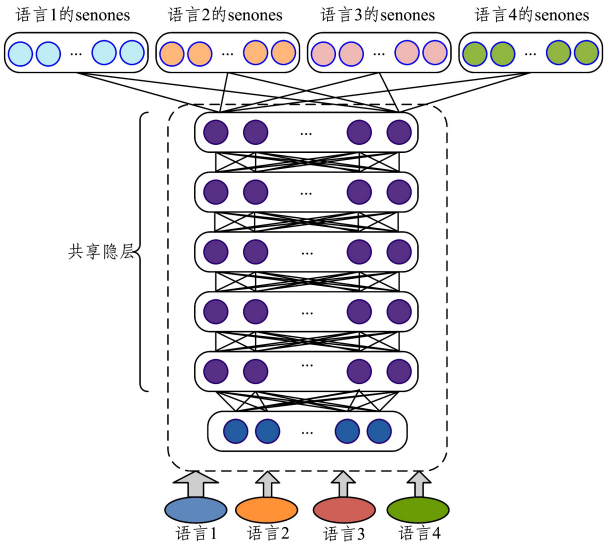

图 1 共享隐层多语言深度神经网络的结构图

Fig. 1 Structure of shared hidden layer multilingual deep neural network

3.2 多语言深度 Bottleneck 神经网络

多语言深度 Bottleneck 神经网络的构建与共享隐层多语言深度神经网络的构建类似,不同之处在于多语言深度 Bottleneck 神经网络包含一层只有少量节点的 Bottleneck 层,而且该 Bottleneck 层通常是线性层。如果多语言深度 Bottleneck 神经网络的 Bottleneck 层处于该深度神经网络较低的层次且仅有少量的目标语言语音数据,如 10 小时的目标语言语音数据,则利用目标语言再进行细调或更新时基本没有效果。

经过多语言深度 Bottleneck 神经网络,通过神经网络前向算法可以为目标语言抽取 Bottleneck 特征。将抽取的目标语言的 Bottleneck 特征与用于抽取 Bottleneck 的特征相拼接,组成新的特征以训练新的深度神经网络。在训练多语言深度 Bottleneck 深度神经网络时,原始的输入特征经过层层的非线性变换和抽象,获得的 Bottleneck 特征对于目标语言的输出具有很好的辨识效果。

已有研究表明[16],基于多语言深度 Bottleneck 神经网络抽取的 Bottleneck 特征较 SHL-MDNN 跨语言迁移学习获得的深度神经网络对目标语言的语音识别更有效。

4 实验

4.1 实验设置

表 1 列出了多个用于训练多语言深度神经网络的带标注的不同语言的语音数据。

表 1 不同语言的语音数据

Table 1 Speech data of different languages

类型	语言	大小/h	区域
源语言	孟加拉语(Bengali)	87	南 亚
	克里奥耳语(Creole)	83	非 洲
	老挝语(Lao)	86	东南亚(带调)
	库尔德语(Kurdish)	51	中 东
	祖鲁语(Zulu)	84	非 洲
	哈萨克语(Kazakh)	54	中 东
	立陶宛语(Lithuanian)	54	东 欧
	瓜拉尼语(Guarani)	55	拉丁美洲
	阿姆哈拉语(Amharic)	56	非 洲
	爪哇语(Javanese)	59	东南亚 太平洋中 南诸岛
	伊博语(Igbo)	56	非 洲
	卢欧语(Dholuo)	54	非 洲
	广东话(Cantonese)	175	东 亚(带调)
	土耳其语(Turkish)	107	中 东
	塔加拉族语(Tagalog)	115	东南亚 太平洋中 南诸岛
	普什图语(Pushto)	111	中 东
	阿萨姆语(Assamese)	74	南 亚
目标语言	蒙古语(Mongolian)	10	中东东亚

在表 1 中,源语言数据用于训练共享隐层多语言深度神经网络和多语言深度 Bottleneck 神经网络。目标语言是带标注的 10 小时蒙语数据。表 1 所列的语音数据都是自然风格的对话语音数据,且这些语言数据以 8000 Hz 的采样率进行

采样后按照 16 位进行量化编码。这些语音对话涉及的领域广泛,如购物、聊天、用餐、看医生等日常对话。

在表 1 中,用于训练共享隐层多语言深度神经网络的 10 种源语言包括孟加拉语、克里奥耳语、老挝语、库尔德语、祖鲁语、哈萨克语、立陶宛语、瓜尼拉语、阿姆哈拉语、爪哇语。用于训练多语言深度 Bottleneck 神经网络的 8 种源语言包括克里奥耳语、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。所有的这些语言数据都是 IARPA Babel 项目数据,可以从 LDC 获得。在选择这些语言数据来训练共享隐层多语言深度神经网络和多语言深度 Bottleneck 神经网络时,考虑了语言自身的特点及其地域分布,如在不同的国家和地区、不同语系等,从而使得训练出的共享隐层多语言深度神经网络或多语言深度 Bottleneck 神经网络具有较好的代表性和较好的泛化性能。表 1 所列的这些语言数据都来自同一个项目,其录音环境、通道等基本一致,避免了由于环境或通道因素的变化对系统性能产生的影响。

在构建共享隐层多语言深度神经网络和多语言 Bottleneck 神经网络时,有 6 个隐层。除 Bottleneck 层含有 42 个隐层单元外,其他每一个隐层含都有 2048 个隐层单元。每种语言大约有 3500～4500 个 senones。

在训练共享隐层多语言深度神经网络和多语言深度 Bottleneck 神经网络时,其语音的强制对齐信息由每种语言相应的 GMM-HMM 系统产生。用开源工具包 Kaldi 训练共享隐层多语言深度神经网络、多语言深度 Bottleneck 神经网络以及每种语言的 GMM-HMM 系统[17]。13 维的梅尔倒谱系数(MFCC)特征以及它们的一阶、二阶差分,共 39 维特征,用于训练每种语言的 GMM-HMM 系统。用基于最大互信息(MMI)的区分度准则来训练每种语言的 GMM-HMM 系统。在训练共享隐层多语言深度神经网络以及多语言深度 Bottleneck 神经网络时,所用的特征是 40 维的 FBank 特征以及 3 维 Kaldi 的 Pitch 特征,共 43 维。基于交叉熵(Cross-entropy)的训练准则用于训练 SHL-MDNN 多语言深度神经网络和多语言深度 Bottleneck 神经网络,并设置初始学习速率为 0.008,且按参数为 0.5 的指数准则对学习速率进行衰减。

在进行跨语言迁移学习,获得目标语言的深度神经网络,然后深度神经网络再按照 sMBR 准则进行序列训练(Sequence Training)。利用多语言深度 Bottleneck 神经网络抽取目标语言的 Bottleneck 特征,并将其与用于抽取 Bottleneck 特征的原始目标语言的 43 维 FBank+Pitch 特征进行拼接获得 85 维特征,然后将该特征用于训练深度神经网络。训练该深度神经网络时,先基于交叉熵进行训练,然后再按照 sMBR 准则进行序列训练(Sequence Training)。在基于交叉熵准则进行训练时,其初始学习速率设置为 0.008,并按参数为 0.5 的指数准则对学习速率进行衰减。

由于仅有 10 小时的目标语言训练数据,仅基于训练数据的脚本不能获得较好的语言模型,因此也不能获得较好的识别效果,表 2 的实验结果也证实了这一点。因此,提出利用网络来获取一定数量的网页,以增强语言模型。实验时,可利用两种不同方法获取一定数量的页面,一种是利用搜索引擎 Bing 和 Google 搜索包含特定词的页面,另一种是利用网络爬虫(如 Scrapy)定向抓取某些站点。实验中,在 2016 年 4 月—5 月期间,主要抓取了中国蒙古语信息网(蒙古文版)、人民网(蒙古文版)、好乐宝博客网等站点。两种方法总共获取了 20 万个不同的页面。页面涉及的内容比较广泛,如广告、新闻、小说、博客等。对抓取的数据进行清洗,最终余下 877 M 文本数据。这些数据用于构建 3-gram 的 Web 语言模型。

在构建资源稀缺的目标语言——蒙语语音识别系统时,所用的带标注语音数据是时长仅为 10 小时的自然口语风格的语音数据。利用这些标注的语音数据构建的 GMM-HMM 包含 2009 个 senones,并且该 GMM-HMM 用于强制对齐训练数据以训练基线深度神经网络模型。这 10 小时的训练标注脚本中包含了 8507 个词。用时长为 10 小时的自然口语风格的蒙语数据评测系统的性能。基于训练脚本构建的 3-gram 语言模型在评测脚本上的困惑度是 154.3。由于抽取的 Web 数据是从大量的包含不同内容的页面获得的,因此它与训练脚本的领域不一致。为了获得领域语言模型,利用 Web 数据训练获得的语言模型与训练脚本获得的语言模型进行插值以获得新的 Web 语言模型,记该语言模型为插值的 Web 语言模型。插值的权值根据评测脚本的困惑度调优。SRILM 工具包用于训练语言模型以及插值权值的调优[18]。在构建 3-gram 语言模型时,词表大小为 23989。插值的语言模型在评测脚本上的困惑度为 127.7。使用词错误率(Word Error Rate,WER)来评测系统的性能。

使用 Linux Ubuntu 14.04 Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz 服务器来训练这些模型。该服务器内存为 265GB,且有 2 块 K20 NVIDIA GPU 卡用于加速深度神经网络的训练。通过开源 Kaldi 工具及相关脚本来完成所有的神经网络模型训练和语音识别实验[17]。

4.2　实验结果

基于目标语言构建的基线系统的识别结果如表 2 所列,该实验结果与文献[12]中的基线系统的结果一致[12]。

表 2　基线系统的识别结果

Table 2　Results of baseline speech recognition system

(单位:%)

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

从表 2 中可以看出:1)由于仅使用时长为 10 小时的训练数据来构建系统,系统性能有较高的词错误率;在利用训练脚本构建的语言模型进行解码的情况下,*WER* 为 73.1%。2)结合Web数据训练的语言模型可以降低其在评测脚本上的困惑度,从而提高识别系统的性能。在基于插值的 Web 语言模型解码的情况下,*WER* 为 71.8%。相较于基于训练脚本语言模型解码的结果,在基于插值的 Web 语言模型解码的情况下,系统错误率有 1.3%的绝对减少。

为了获得较好的识别效果，在进行跨语言的迁移学习时，对基于 10 种语言训练获得的多语言深度神经网络不同数量的隐层和输出层进行了细调。图 2 给出了细调不同层数的隐层和输出层的深度神经网络获得的识别系统的性能。其中深度神经网络仅采用交叉熵准则训练，且仅利用基于训练脚本获得的语言模型解码的结果。在细调不同层数的隐层和输出层时，设置初始的学习速率为 0.002。

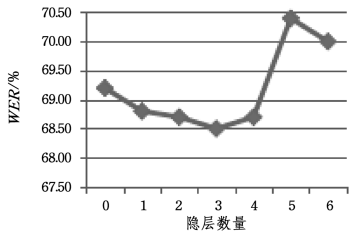

图 2 细调不同隐层个数的神经网络在评测集上获得的识别结果

Fig. 2 Recognition results of neural network on evaluation data set by fine-tuning different hidden layers

由图 2 可知，识别系统在细调 3 个隐层和 1 个输出层时获得了最佳的识别结果。

基于 SHL-MDNN 多语言深度神经网络跨语言迁移学习获得的深度神经网络识别系统的性能如表 3 所列。在表 3 中，所有深度神经网络都是利用 sMBR 准则进行序列训练(Sequence Training)而获得的。

表 3 基于多语言深度神经网络迁移识别系统的识别结果

Table 3 Recognition results of transferred recognition system based on multilingual deep neural network

系统		*WER*/%	
名称	系统描述	训练脚本语言模型	Web 语言模型
Multilingual-10-DNN	基于 10 种语言的深度神经网络迁移获得的神经网络	65.1	62.5
Multilingual-8-BNF	结合基于 8 种语言的深度 Bottleneck 抽取特征训练获得的深度神经网络	63.3	61.1

从表 3 中可以看到：无论是利用 SHL-MDNN 多语言深度神经网络进行跨语言迁移学习，还是利用多语言深度 Bottleneck 神经网络提取 Bottleneck 特征，都可以使资源稀缺语言识别系统的性能得到提高。在利用插值的 Web 语言模型解码的情况下，*WER* 较表 2 中的基线系统有9.3%～10.7%的绝对减少。对比基于 SHL-MDNN 多语言深度神经网络迁移学习获得的神经网络和基于抽取的多语言深度 Bottleneck 神经网络获得的深度神经网络，可以看到：结合多语言深度 Bottleneck 神经网络提取的 Bottleneck 特征进行训练获得的深度神经网络的识别效果更好。

文献[3]的研究表明，使用与目标语言相似的多语言数据训练获得的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助于目标语言识别系统性能的提高。基于该项研究结果，利用第 2 节中提出的基于长短时记忆递归神经网络语言辨识方法的多语言语音数据选择方法，10 种语言(哈萨克语、克里奥耳语、阿萨姆语、普什图语、立陶宛语、库尔德语、瓜尼拉语、阿姆哈拉语、老挝语、孟加拉语)的语音数据的句子被选出用于训练 SHL-MDNN 多语言深度神经网络和多语言深度 Bottleneck 神经网络。将蒙古语作为目标语言进行跨语言迁移学习，并基于获得目标语言的深度神经网络模型进行解码，实验结果如表 4 所列。

表 4 由 LSTM-RNN LID 选择数据训练 DNN 迁移获得的识别系统的结果

Table 4 Results of transferred deep neural network trained by LSTM-RNN through data training

系统		*WER*/%	
名称	系统描述	训练脚本语言模型	Web 语言模型
LSTM-RNN-LID-10-DNN	基于语言辨识方法选择出的 10 种语言的句子进行迁移学习获得的神经网络	63.8	61.3
LSTM-RNN-LID-10-BNF	基于语言辨识方法选择出的 10 种语言的句子训练获得的深度神经网络	62.9	60.4

对比表 3 可以看到，相比用随机方法选择的 10 种语言数据训练获得 SHL-MDNN 多语言神经网络和基于 Bottleneck 特征训练获得深度神经网络的性能，基于语言辨识方法选择出的 10 种语言的句子训练的深度神经网络或者基于 Bottleneck 抽取器抽取的 Bottleneck 训练获得的神经网络(LSTM-RNN-LID-0-DNN 和 LSTM-RNN-10-BNF)的性能都有不同程度的提高。在基于训练脚本的训练获得语言模型进行解码时，错误率分别有 1.3%和 0.4%的绝对减少；在基于插值的 Web 语言模型进行解码时，错误率分别有 1.2%和 0.7%的绝对减少。

相比表 2 中的基线系统的性能，基于语言辨识方法选择出的 10 种语言的句子训练的深度神经网络的性能或者基于 Bottleneck 抽取器抽取的 Bottleneck 训练获得的神经网络(LSTM-RNN-LID-0-DNN 和 LSTM-RNN-10-BNF)的性能都有不同程度的提高。在基于训练脚本的训练获得语言模型进行解码时，错误率分别有 9.3%和 10.2%的绝对减少；在基于插值的 Web 语言模型进行解码时，错误率分别有 10.5%和 11.4%的绝对减少。

结束语 构建当前最先进的语音识别系统需要大量带标注的语音数据，因此，在缺少大规模带标注的蒙语语音数据的情况下构建最先进的蒙语语音识别系统较困难。多语言和跨语言信息的有效利用可以提高识别系统的声学建模能力，从而有助于提高语音识别系统的性能。而基于长短时递归神经网络语言辨识方法的数据选择可以选择出对目标语言更有利的多语言语音数据，从而训练与目标语言相关的多语言深度神经网络，有助于提高资源稀缺目标语言语音识别系统的性能。与从不同的多语言获得的神经网络以及基线系统相比，基于本文提出的方法选出的多语言语音数据训练深度神经网

络迁移学习获得的深度神经网络的性能在不同的语言模型解码条件下,错误率有不同程度的减少。带标注的目标语言语音数据的缺乏使得识别系统的性能不能达到最好。未来将探讨增加数据或采用半监督的机器学习方法弥补数据缺乏的不足,以期能够提高识别系统的性能。

参 考 文 献

[1] Ethnologue. Ethnologue languages of the world [OL]. http://www.ethnologue.com.

[2] BRANDING C C. Summer Institute for Linguistics Ethnologue Survey1999 [OL]. https://afrobranding.wordpress.com/tag/summer-institute-for-linguistics-sil-ethnologue-survey.

[3] ZHANG Y,CHUANGSUWANICH E,GLASS J. Language ID-based Training of Multilingual Stacked Bottleneck Features [C]// Proceedings of INTERSPEECH. Singapore: IEEE Press, 2014: 1-5.

[4] KNILL K M,GALES M J F,RATH S P,et al. Investigation of Multilingual Deep Neural Networks for Spoken Term Detection [C]// Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop. Olomouc IEEE Press,2013:138-143.

[5] GHOSHAL A,SWIETOJANSKI P,RENALS S. Multilingual Training of Deep Neural Networks [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Proceessing. Vancouver:IEEE Press,2013:7319-7323.

[6] HUANG J T,LI J,YU D,et al. Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Proceessing. Vancouver: IEEE Press,2013:7304-7308.

[7] VU N T,IMSENG D ,POVEY D,et al. Multilingual Deep Neural Network based Acoustic Modeling for Rapid Language Adaptation[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Proceessing. Florence: IEEE Press,2014:7639-7643.

[8] CUI J,KINGSBURY B,RAMABHADRAN B,et al. Multilingual Representation for Low-resource Speech Recognition and Keyword Search [C]// Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop. Scottsdale: IEEE Press,2015:259-266.

[9] SIBO T,PHILIP N G,HERVE B. An Investigation of Deep Neural Networks for Multilingual Speech Recognition and Adaptation [C]// Proceedings of INTERSPEECH. Stockholm: IEEE Press,2017:714-718.

[10] LU Y,LU F,SEHGAL S,et al. Multitask Learning in Connectionist Speech Recognition [C]// Proceedings of Australian International Conference on Speech Science and Technology. Sydney:IEEE Press,2004:312-315.

[11] 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]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Proceessing. Florence: IEEE Press, 2014: 5592-5596.

[12] ZHANG A Y,NI C J. Research on Low-resource Mongolian Speech Recognition[J]. Computer Science, 2017, 44(10): 318-322. (in Chinese)
张爱英,倪崇嘉. 资源稀缺蒙语语音识别研究[J]. 计算机科学, 2017,44(10):318-322.

[13] NI C,LEUNG C C,WANG L,et al. Efficient Methods to Train Multilingual Bottleneck Feature Extractors for Low Resource Keyword Search[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Proceessing. New Orleans:IEEE Press,2017:5650-5654.

[14] NI C,WANG L,LEUNG C C,et al. Rapid Update of Multilingual Deep Neural Network for Low-Resource Keyword Search [C]// Proceedings of INTERSPEECH. San Francisco: IEEE Press,2016:3698-3702.

[15] GONZALEZ-DOMINGUEZ J, LOPEZ-MORENO I,SAK H,et al. Automatic Language Identification Using Long Short-Term Memory Recurrent Neural Networks [C] // Proceedings of NTERSPEECH. Singapore:IEEE Press,2014:2155-2159.

[16] 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] // Proceedings of INTERSPEECH. Dresden:IEEE Press, 2015:2132-2136.

[17] POVEY D,GHOSHAL A,BOULIANNE G,et al. The Kaldi Speech Recognition Toolkit [C]// Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Hawaii:IEEE Press,2011:1-4.

[18] STOLCKE A. SRILM-An Extensible Language Modeling Toolkit [C]// Proceedings of International Conference on Spoken Language Processing. Denver:IEEE Press,2002:901-904.