DOI：10.7652/xjtuxb201710020

采用长短时记忆网络的低资源语音识别方法

舒帆[1]，屈丹[1]，张文林[1]，周利莉[1]，郭武[2]
（1.解放军信息工程大学信息系统工程学院，450002，郑州；
2.中国科学技术大学信息科学技术学院，230026，合肥）

摘要：针对低资源环境下由于标注训练数据不足、造成语音识别系统识别率急剧下降的问题，提出一种采用长短时记忆网络的低资源语音识别（LSTM-LRASR）方法。该方法采用长短时记忆网络构建声学模型，从特征提取、数据扩展及模型优化3个方面提高低资源语音识别性能。在特征提取方面，提取语言无关的高层稳健特征参数，降低声学模型对训练数据的依赖；在数据扩展方面，对已有标注数据进行语速扰动，对无标注数据进行自动识别，从而自动获取更多标注数据；在模型优化方面，通过序贯区分性训练技术提高模型对易混淆音素的区分能力，利用最小风险贝叶斯解码对多个系统进行融合，进一步提高识别性能。对OpenKWS16评测数据的实验结果表明，采用LSTM-LRASR方法搭建的低资源语音识别系统的词错率相对基线系统下降了29.9%，所有查询词的查询项权重代价提升了60.3%。

关键词：语音识别；低资源；长短时记忆；神经网络
中图分类号：TN912.34 **文献标志码**：A **文章编号**：0253-987X(2017)10-0120-08

A Speech Recognition Method Using Long Short-Term Memory Network in Low Resources

SHU Fan[1], QU Dan[1], ZHANG Wenlin[1], ZHOU Lili[1], GUO Wu[2]
(1. Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou 450002, China;
2. Institute of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: A speech recognition method using long short-term memory network in low resources (LSTM-LRASR method) is proposed to solve the problem that the recognition rate of an auto speech recognition system is declining due to the lack of transcripted training data in low resource environments. The method uses long short-term memory network to construct an acoustic model, and improves the low resource speech recognition performance from three aspects. These are feature extraction, data augmentation and model optimization. The feature extraction extracts language-independent high-level robustness parameters to reduce the dependence of acoustic model on training data. The data augmentation processes the transcripted data by speed perturbation, while the untranscripted data is recognized automatically, so that more transcripted data are created. The model optimization uses the sequential discriminating training technique to improve the ability of distinguishing phonemes, and the minimum Bayes-risk decoding is used to combine multiple systems and to further improve the recognition performance. The experimental

收稿日期：2017-01-17。 作者简介：舒帆(1992—)，男，硕士生；屈丹(通信作者)，女，副教授。 基金项目：国家自然科学基金资助项目(61673395,61403415,61302107)；河南省自然科学基金资助项目(162300410331)。
网络出版时间：2017-07-14 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20170714.2044.016.html

results on the OpenKWS16 evaluation database show that the word error rate of the low resource speech recognition system built by the proposed LSTM-LRASR method is 29.9% lower than that of the baseline system, and the actual value weighted value increases by 60.3%.

Keywords: speech recognition; low resource; long short-term memory; neural network

目前的自动语音识别(auto speech recognition, ASR)技术依赖于大量的数据资源，在低资源环境下，语音识别系统的识别率将明显下降。世界上总共约有 6 900 种语言，仅有为数不多的几种语言(如英语、汉语普通话等)具有充足的数据资源，大部分语言都是低资源的。随着经济全球化的深入发展，语音识别技术的应用不再局限于英语、汉语普通话等高资源语言，如何在低资源环境下构建高性能的语音识别系统已成为国际上的研究热点与难点问题。2011 年，美国情报先进研究计划署提出 Babel 计划[1]。该计划为期 5 年，其研究目标是利用有限数据资源快速开发针对任意未知语言且性能稳健的语音识别系统。众多国内外研究机构如 BBN 公司、IBM 公司、卡内基梅隆大学、约翰霍普斯金大学、布尔诺科技大学、清华大学、中国科技大学等都参与了该计划，并取得了一定的研究成果[2-7]。2016 年，本团队参加了 Babel 计划中公开关键词检索(open keyword search, OpenKWS)评测[8]项目，所搭建的语音识别系统的连续语音识别性能在所有团队中名列第 4，在国内排名第 1。

针对低资源环境下由于标注训练数据不足，造成语音识别系统识别率急剧下降的问题，本文提出一种采用长短时记忆(long short-term memory, LSTM)网络的低资源语音识别(LSTM-LRASR)方法。该方法在 LSTM 声学模型的基础上，先根据多语言数据共享的思想，得到多语言共享的瓶颈特征，再采用数据扩展技术和半监督训练方法获取更多标注训练数据，然后通过序贯区分性训练提升声学模型区分能力，最后采用系统融合充分利用不同改进模型各自优势。采用本文方法在 OpenKWS16 评测数据的连续语音识别和关键词检索实验中进行测试，实验结果表明，本文方法在低资源语言上具有较好的适应性，能有效提升低资源语音识别系统的识别性能。

1　采用 LSTM 的声学建模

语音信号是一种复杂的时变信号，在不同的时间范围内具有复杂多变的相关性。传统的高斯混合模型(Gaussian mixture model, GMM)、子空间高斯混合模型(subspace Gaussian mixture model, SGMM)和深层神经网络(deep neural network, DNN)受到固定窗长限制，只能对窗内有限时间数据进行建模，无法解决具有不同说话速率或长距离相依的问题，而传统循环神经网络(recurrent neural network, RNN)虽然能够利用过去时间信息，兼顾不同时间范围的数据相关性，但它存在的梯度消失问题难以解决。LSTM 网络用一种新的结构单元代替传统 RNN 中的神经元，避免了传统 RNN 模型的梯度消失和梯度爆炸问题，能够对长时语音信息进行建模。

1.1　LSTM 网络结构

LSTM 基本结构单元又称作记忆块，它代替传统神经元单元组成 LSTM 网络。记忆块由 1 个中心节点和 3 个门控单元组成。3 个门控单元与中心节点分别相连，分别称作输入门、输出门和遗忘门，它们控制信息流入，中心节点称作记忆细胞，用于存储当前网络状态。在前向传递中，输入门控制输入到记忆细胞的信息流，输出门控制记忆细胞到网络其他结构单元的信息流；在后向传递中，输入门控制误差流出记忆细胞，输出门控制误差流入记忆细胞。遗忘门控制记忆细胞内部的循环状态，决定记忆细胞中信息的取舍(遗忘)。LSTM 的这种门控机制是让信息选择性通过，使记忆细胞具有保存长距离相依信息的能力，并在训练过程中防止内部梯度受外部干扰。每个记忆细胞都存在一个自循环连接线性单元，称为恒定误差传送带，使误差在内部以恒定值传播，避免了梯度消失和梯度爆炸问题。

已知输入为 x_t，LSTM 记忆块的输出 y_t 按照以下公式迭代计算

$$i_t = \sigma(W_{ix}x_t + W_{im}y_{t-1} + W_{ic}c_t + b_i) \quad (1)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}y_{t-1} + W_{fc}c_t + b_f) \quad (2)$$

$$o_t = \sigma(W_{ox}x_t + W_{om}y_{t-1} + W_{oc}c_t + b_o) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \quad (4)$$

$$y_t = o_t h(c_t) \quad (5)$$

式中：$t=1,2,\cdots,T$；i_t、f_t、o_t 和 c_t 分别是输入门、遗忘门、输出门和记忆细胞的输出；W_{ix}、W_{im} 和 W_{ic} 分别表示网络输入、上时刻输出和记忆细胞到输入门的权重；W_{fx}、W_{fm} 和 W_{fc} 分别为网络输入、上一时刻

输出以及记忆细胞到遗忘门的权重；W_{ox}、W_{om} 和 W_{oc} 分别是其他单元到输出门的权重；b_i、b_o、b_f 和 b_c 分别是输入门、输出门、遗忘门和记忆细胞的偏移量；$\sigma(\cdot)$ 是对数 sigmoid 函数；$g(\cdot)$ 和 $h(\cdot)$ 分别是记忆细胞的输入和输出激活函数，通常取双曲正切函数 tanh(·)。

LSTM 基本结构单元如图 1 所示。

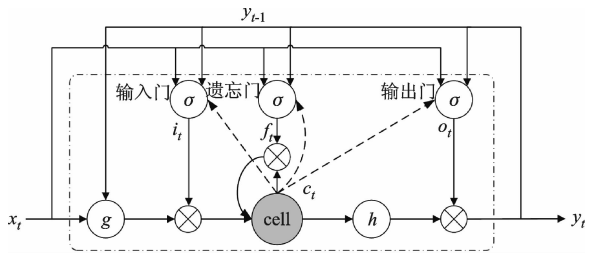

图 1 LSTM 基本结构单元

1.2 LSTM 语音声学模型

为了实现更好的声学建模，将 LSTM 网络与传统隐马尔可夫模型(hidden Markov model，HMM)结合，构造了声学模型 LSTM-HMM[9]。图 2 给出了基于 LSTM 声学模型的语音识别系统框图。系统利用 LSTM 进行声学建模，代替传统的高斯混合模型 GMM 和子空间高斯混合模型 SGMM 进行状态输出概率的计算，并利用 HMM 计算各状态之间的转移概率；然后，在语言模型和发音字典的共同作用下解码得到识别结果。

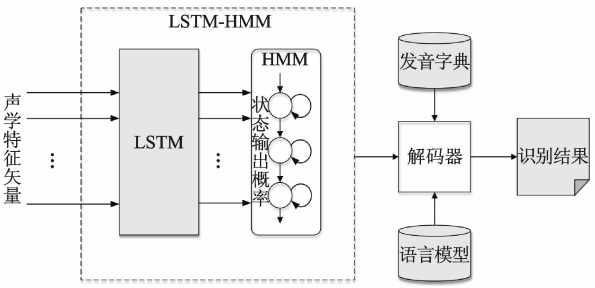

图 2 基于 LSTM 声学模型的语音识别系统框图

2 采用 LSTM 网络的低资源语音识别

文本将 LSTM 网络与多种技术结合，提出采用 LSTM 网络的低资源语音识别(LSTM-LRASR)方法。该方法采用 LSTM 网络构建声学模型，从特征提取、数据扩展及模型优化等方面提高低资源语音识别性能。

2.1 多语言瓶颈特征

多语言语音技术是低资源语言语音识别的一项关键技术。通过多语言语音技术可以借用现有的其他语言数据资源，弥补低资源语言训练数据不足的问题。其中，多语言瓶颈特征(multilingual bottleneck feature，MBNF)技术[10]是目前最常用的多语言语音技术。

多语言瓶颈特征通常利用一种具有特殊的神经网络提取。首先，这种网络的隐含层中，有个隐含层的神经元数较其他隐含层少得多，使得整个网络呈瓶状，因此将这个特殊的隐含层形象地称为瓶颈(bottleneck，BN)层，MBNF 特征便是从该 BN 层提取获得。其次，神经网络的参数由多种语言数据训练得到。由于不同语言通常有各自的发音特点，其音素、音节等子词单元不尽相同，因此在神经网络进行多语言训练时一般要用到多任务学习方法，即在训练过程中，不同语言数据共享隐含层，但在神经网络输出端有其对应的 softmax 层。

提取 MBNF 的一般流程如图 3 所示。图中给出的神经网络有 5 个隐含层，BN 层控制着整个网络从输入到输出之间的信息流。由于 BN 层的神经元个数相对其他隐含层很小，神经网络训练将会使得 softmax 层输出的分类信息在 BN 层内尽可能压缩，实现有效的非线性降维。训练完成后，取 BN 层的先行输出为 MBNF，可直接用于声学模型，也可以先经过一定后处理，如说话人自适应等，再用于声学模型。

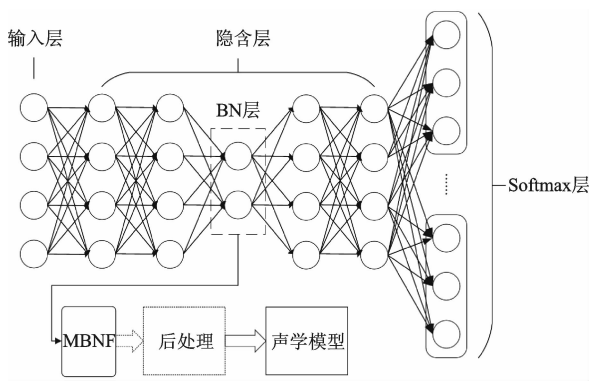

图 3 MBNF 一般提取流程

BN 层可以位于网络中的任何位置，图 3 中采用第 3 个隐含层作为 BN 层。由于神经网络中每个隐含层的状态输出都是输入矢量的一个新的特征表示，且层次越深表征越抽象，因此通常把 BN 层放在网络靠后的位置。

2.2 半监督训练

常规的有监督方法只能使用有标注数据进行模型训练，但是对语音数据进行标注是一件费时费力的工作，尤其对于低资源语言，往往有标注数据很

少，大部分语音数据没有文本标注。针对无标注数据，可使用半监督训练方法[11]。

半监督训练是一种自学习的训练方法。半监督训练过程先用有标注数据训练 ASR 系统，该系统被称为种子系统，随后用种子系统对无标注语音数据进行解码。假设系统对无标注数据的解码是准确的，便可以将解码结果当作为标注数据的文本标注，原先的无标注数据也可以当作有标注数据使用，加入训练集，用于之后的模型训练。但实际上，这样得到的标注是不完全准确的，因此有时还会对识别后的无标注数据进行数据筛选再加入训练集。对新生成的标注（即解码结果）计算置信度得分，筛选对应得分较高无标注数据加入训练集。常用的置信度得分为词或句子的后验概率。

2.3　语速扰动

在保留原有标注的情况下，将原有数据进行变换后再加入训练数据已被证实为神经网络训练时扩充数据的一种有效方法。为了弥补低资源环境下训练数据不足的问题，本文使用了语速扰动方法对原有训练数据进行了数据扩充[12]。

语速扰动是语音信号的一种变换方法，该方法产生一个规整的时间信号。给定语音信号 $x(t)$ 和规整因子 α，则扰动后的语音信号为

$$x'(t) = x(\alpha t) \tag{6}$$

式(6)的傅里叶变换为

$$X'(\omega) = \alpha^{-1} X(\alpha^{-1}\omega) \tag{7}$$

式中：$X'(\omega)$ 是语音信号 $x'(t)$ 的傅里叶变换。语速扰动实质上是对语音信号在时域进行线性拉伸或压缩。信号长度改变，信号分帧情况将会受到影响。式(7)表明，扰动后的语音信号在频域也会按照一定比例发生变化。当 $0<\alpha<1$，扰动后语音信号的语速变慢，能量向低频移动；反之，若 $\alpha>1$，扰动后信号的语速加快，能量转向高频。由于 Mel 频率与频率有着对数关系，Mel 频谱也会相应发生偏移。在时域和频域变化的共同作用下，信号经特征提取后会和原信号呈现差异性，因此可以将原始训练数据经语速扰动后再加入训练集，实现数据扩充。

2.4　序贯区分性训练

序贯区分性训练是对模型进一步调优的重要手段[13]。通常，神经网络基于交叉熵（cross-entropy，CE）准则进行训练，该准则没有考虑语音帧的序列信息。为了使系统达到更好的性能，在神经网络训练完成后，通常还会基于一些其他具有序列区分能力的准则进行调优，此时的训练被称为序贯区分性训练。序贯区分性训练可采用的准则有很多种，如最大互信息准则、最小分类错误准则、最小音素错误准则、状态级最小风险贝叶斯（state-level minimum Bayes risk，sMBR）准则等。本文根据文献[14-15]中的建议，采用 sMBR 准则进行区分性训练。

与关注语音帧分类能力的 CE 准则不同，sMBR 准则直接针对语音信号的状态序列，准则目标是使识别中状态标签的错误率达到最小，sMBR 准则的目标函数为

$$F_{\mathrm{sMBR}} = \sum_{u} \frac{\sum_{\boldsymbol{W}} p(\boldsymbol{O}_u \mid \boldsymbol{S}_u)^{\kappa} P(\boldsymbol{W}) A(\boldsymbol{W}, \boldsymbol{W}_u)}{\sum_{\boldsymbol{W}'} p(\boldsymbol{O}_u \mid \boldsymbol{S}_u)^{\kappa} P(\boldsymbol{W}')} \tag{8}$$

式中：$\boldsymbol{O}_u = \{\boldsymbol{o}_{u1}, \boldsymbol{o}_{u2}, \cdots, \boldsymbol{o}_{uT_u}\}$ 是语句 u 的所有观测向量组成的序列，$\boldsymbol{o}_{ut}$ 为第 t 个语音帧的观测向量（$t=1,2,\cdots,T_u$），T_u 是该观测序列的长度；$\boldsymbol{W}_u$ 是 u 对应标注的词序列；$\boldsymbol{S}_u$ 是 $\boldsymbol{W}_u$ 对应的状态序列；κ 是声学缩放因子；$A(\boldsymbol{W},\boldsymbol{W}_u)$ 描述标注词序列 $\boldsymbol{W}_u$ 和词序列 $\boldsymbol{W}$ 间的准确度，一般选取 $\boldsymbol{W}$ 和 $\boldsymbol{W}_u$ 之间对应正确的状态标签个数；$P(\boldsymbol{W})$ 是语言模型概率；$p(\boldsymbol{O}_u \mid \boldsymbol{S}_u$ 是声学似然度。序贯区分性训练调整声学模型（神经网络）参数，使目标函数达到最小。

2.5　系统融合

系统融合是提升系统性能的重要手段，它主要通过利用不同系统间的互补性得到更好的识别结果[16]。本文对搭建的多个系统在两个层面分别进行了系统融合：一是针对连续语音识别，利用最小风险贝叶斯（minimum Bayes risk，MBR）解码方法对多个系统生成的同一语句的多个词格进行联合解码[17]，得到新的识别结果，计算词错率；二是关键词检索，对独立系统得到的关键词检索列表进行重打分，获得新的关键词检索列表，计算实际查询项权重代价。

设有 N 个独立系统，融合时各个系统的权重用 λ_i 表示（$i=1,2,\cdots,N$；$\sum_{i=1}^{N}\lambda_i = 1$）。MBR 解码的目标函数为

$$F_{\mathrm{MBRd}} = \sum_{i=1}^{N} \lambda_i \frac{\sum_{\boldsymbol{W}} p_i(\boldsymbol{O}_u \mid \boldsymbol{S}_u)^{\kappa} P(\boldsymbol{W}) A(\boldsymbol{W}, \boldsymbol{W}_u^*)}{\sum_{\boldsymbol{W}'} p_i(\boldsymbol{O}_u \mid \boldsymbol{S}_u)^{\kappa} P(\boldsymbol{W}')} \tag{9}$$

式中：$\boldsymbol{W}_u^*$ 表示语句 u 的识别词序列，其他符号定义与式(8)相同。MBR 解码过程找到最佳词序列 $\boldsymbol{W}_u^*$ 使目标函数最小。在式(9)的约束下，系统融合输出

词序列必将出现在至少一个独立系统中[17]。

关键词检索列表重打分结果由各个独立系统关键词检索列表得分加权计算得到。各系统权重仍用 λ_i 表示,独立系统中某个关键词检索项的置信度得分记为 s_i,系统融合得分

$$s_{\text{comb}} = \left(\sum_{i=1}^{N}\lambda_i s_i^{\alpha}\right)^{-\alpha} \tag{10}$$

式中:α 是常数,本文取 $\alpha=0.1,0.2,\cdots,0.9$ 分别进行了实验,最终选择使系统融合得分最高的 α 值。

3 实验及结果分析

为了衡量本文方法的性能,采用 OpenKWS16 评测数据进行实验,以连续语音识别的词错率和关键词检索的实际查询项权重代价作为性能评价指标进行测试[8]。

3.1 实验设置

在 OpenKWS16 评测语料上进行实验,针对格鲁吉亚语搭建 ASR 系统。格鲁吉亚语的完整语言包(full language pack,FLP)的训练集中有约 80 h 的对话音频数据,其中只有 50%的音频数据具有语料标注,训练可用于模型训练。FLP 中还有 10 h 的开发集数据,开发集数据只用来给各团队测试所搭建系统的性能,不可用于模型训练。除了格鲁吉亚语资源,评测还允许团队利用多语言技术使用其他语言资源。Babel 计划发布数据资源如表 1 所示。

表 1 Babel 计划发布数据资源表

语言	数据量/h	语言	数据量/h
粤语	140.71	宿务语	41.01
普什图语	77.30	哈萨克语	39.59
土耳其语	76.38	泰卢固语	41.47
塔加拉族语	83.78	立陶宛语	42.08
越南语	87.15	斯瓦希利语	44.01
阿萨姆语	60.26	瓜拉尼语	42.63
孟加拉语	61.12	伊博语	43.65
海地克里奥尔语	66.51	阿姆哈拉语	43.15
老挝语	65.19	蒙古语	45.71
祖鲁族语	61.32	爪哇语	45.14
泰米尔语	64.04	卢欧语	41.32
库尔曼吉语	41.70	格鲁吉亚语	43.62
巴布亚皮钦语	39.04		

注:数据量是指语言包中标注训练音频的总时长。

基于 Kaldi 工具包搭建格鲁吉亚语 ASR 系统[18],使用标注语料训练得到 3-gram 语言模型,通过查询格鲁吉亚语发音规则建立 G2P 转换生成发音字典,使用 LSTM-HMM 模型搭建声学模型基线系统。

ASR 系统将语音识别成词格形式,在词格的基础上进行解码,转换成文本,实现连续语音识别;在连续语音识别的基础上,系统再进行关键词检索。本文采用加权有限状态转换器进行关键词检索[19]。通常关键词检索列表中存在一些在系统词汇表中未出现的词,这些词被称为集外词;反之,在词汇表中出现的被称为集内词。为了更好地验证模型关键词检索性能,分别单独计算了各系统集内词和集外词的实际查询项权重代价。

3.2 基线系统

采用 LSTM-HMM 作为基线系统声学模型。为了验证 LSTM 在声学建模中的性能,还采用传统的 GMM-HMM 和 SGMM-HMM 声学建模进行对比实验。HMM 有 3 个状态,拓扑结构从左至右无跨越,在最大似然准则下训练得到。由于本文所有系统都使用相同配置的 HMM,为了简化描述,以下分别将 LSTM-HMM、GMM-HMM 和 SGMM-HMM 表示为 LSTM、GMM 和 SGMM。

GMM 声学模型采用 13 维感知线性预测系数加 3 维基音(Pitch)特征[20]作为声学特征,连续 9 帧拼接后用线性判别式分析将维数降到 40 维,然后采用特征空间最大似然线性回归进行说话人自适应训练,再用增强型最大互信息准则做参数调整。Pitch 特征由归一化基音频率、浊音概率及其差分基音频率组成[20]。模型中共有 4 815 个状态和 75 084 个高斯混元。

SGMM 声学模型[21]与 GMM 使用相同的声学特征、说话人自适应方法和区分性准则,而 SGMM 包含 8 930 个状态,每个状态有 800 个高斯混元,子状态数为 80 044。

LSTM 网络输入的声学特征为 40 维滤波器组(Filter Bank,FBank)特征加 3 维 Pitch 以及 100 维 iVector,目标标签有 5 帧时延。本文使用的 iVector 提取器输入为 40 维梅尔频率倒谱系数加 3 维 Pitch。LSTM 网络有 3 个 LSTM 层,每层有 1 024 个记忆块单元,每个 LSTM 层后有 1 个 256 维的线性投影层。

表 2 为 3 种基线系统对开发集的识别性能。本文实验中的关键词列表共 4 469 个关键词,其中集内词 3 903 个、集外词 566 个。从表 2 结果可见,LSTM 声学模型并没有达到理想的效果,除了集外

词的关键词检索性能相对较好外，其他性能都不如传统的 GMM，其主要原因是低资源语言训练数据不足，神经网络训练不充分。

表 2　3 种基线系统的识别性能

声学模型	词错率/%	实际查询项权重代价		
		集内词	集外词	所有查询词
GMM	60.2	0.353 0	0.107 2	0.330 1
SGMM	**52.7**	**0.376 4**	0.122 7	**0.386 6**
LSTM	61.3	0.335 1	**0.157 1**	0.318 5

注：黑体数据表示同一指标下的最佳性能。

3.3　LSTM-LRASR 系统

为降低声学模型对训练数据的依赖性，本文通过 MBNF 技术提取更为稳健的高层声学特征。将格鲁吉亚语外的其他 24 种语言数据都用来训练 BN 网络，输入特征为 40 维 Fbank 加 3 维 Pitch 以及 100 维 iVector，训练采用多任务学习的方式，每种语言对应一个任务。BN 网络使用 p 范数（p-norm）网络结构，网络有 8 个隐含层，每个隐含层后都有若干个 p-norm 单元，p-norm 单元以上层连续 10 个节点输出作为输入向量，计算向量的 p 范数（p=2）。输出第 7 个隐含层为 BN 层，有 1 280 个节点，后面 128 个 p-norm 单元的输出即 MBNF。其他隐含层节点数为4 000，对应有 400 个 p-norm 单元。使用 MBNF 作为声学特征，重新训练模型，系统性能如表 3 所示。比较表 3 和表 2 中的数据可以发现，使用 MBNF 作为声学特征后，LSTM-HMM 系统得到较为明显的提升，词错率下降了 19.7%，总体实际查询项权重代价提升了 36.1%，而 GMM 和 SGMM 性能只是略有提升，LSTM 的连续语音识别性能和关键词检索性能均已超越了 GMM 和 SGMM 模型。

表 3　使用 MBNF 后的 LSTM 声学模型系统性能

声学模型	词错率/%	实际查询项权重代价		
		集内词	集外词	所有查询词
GMM＋M	61.7	0.356 5	0.138 9	0.336 2
SGMM＋M	51.2	0.406 5	0.143 8	0.419 7
LSTM＋M	**49.2**	**0.455 7**	**0.216 7**	**0.433 5**

注："M"表示声学建模使用了多语言瓶颈特征。

对于无标注数据，将半监督训练方法应用到模型训练中，先使用种子系统对这些数据进行识别解码，识别结果将被当作语料标注和这些无标注数据一起加入训练集中。为了提高生成标注文本的准确性，将 SGMM、LSTM、SGMM＋M 和 LSTM＋M 这 4 个 ASR 系统都作为种子系统，对无标注语音数据分别识别成词格形式，然后再将 4 个词格合并。新词格中的置信度得分为 4 个词格中对应置信度得分的线性加权相加，4 个词格的权重都设定为 0.25。最后对新词格解码的结果才作为无标注语料的标注文本。本次实验中，共有无标注数据 39.25 h，与有标注数据量相当，半监督训练使训练集数据量扩大为约原来的 2 倍。表 4 为使用半监督训练方法将无标注数据应用于声学模型训练后系统的词错率和实际查询项权重代价。由表 4 可以看到：半监督训练虽然使用了不完全准确的数据标注用于声学模型训练，但是对系统性能仍很有帮助；相比基线 LSTM 系统，半监督训练系统的词错率下降了 16.6%，总体实际查询项权重代价提升了 32.3%；当 MBNF 和半监督训练同时使用时系统的词错率进一步下降了 5.1%，实际查询项权重代价也得到进一步提升。

表 4　半监督训练模型系统性能

声学模型	词错率/%	实际查询项权重代价		
		集内词	集外词	所有查询词
LSTM＋S	51.1	0.447 4	0.166 7	0.421 3
LSTM＋M＋S	**48.5**	**0.458 0**	**0.221 5**	**0.435 8**

注："S"表示声学建模使用了半监督训练方法。

表 5 给出了数据扩充后几种声学模型的系统性能。文本使用 sox 工具对训练数据重采样，得到 2 组新数据，2 组数据对应的语速分别为原来的 0.9 和 1.4 倍。将原始训练数据与 2 组新数据合并，得到的数据量是原来的 3 倍。与基线 LSTM 系统比较，数据扩展后，词错率降低了 6.4%，总体实际查询项权重代价提升了 8.3%，与半监督训练相比，虽然增加的数据量较多，但并没有得到更好的效果。在使用 MBNF 作为声学特征后，LSTM＋M＋A 的 2 个性能指标均优于 LSTM＋M＋S。造成这种现象的原因主要有两点：①基于数据扰动增加的数据

表 5　数据扩充后的 LSTM 声学模型系统性能

声学模型	词错率/%	实际查询项权重代价		
		集内词	集外词	所有查询词
LSTM＋A	57.4	0.366 9	0.132 8	0.344 9
LSTM＋S＋A	57.3	0.376 5	0.130 7	0.353 6
LSTM＋M＋A	46.7	0.426 1	0.165 9	0.441 0
LSTM＋M＋S＋A	**46.2**	**0.435 4**	**0.223 9**	**0.456 1**

注："A"表示声学建模使用了基于语速扰动的数据扩充方法。

量虽然更多,但是扩展的数据是由原始数据变换得到,它们具有的差异性不如半监督训练中使用的无标注数据;②如果使用表征力更强的特征,能更好地发掘出数据间的差异性,进而提升训练效果,使数据量的优势体现出来。

序贯区分性训练在现有 ASR 系统基础上进行,对神经网络参数进一步调优。序贯区分性训练使用 2.4 节中介绍的 sMBR 准则。由于语料标注是序贯区分性训练中的关键因素,而半监督训练生成的系统并不完全准确,这对序贯区分性训练的效果将产生巨大影响,因此本文并未对半监督训练系统进行后续的序贯区分性训练。表 6 为序贯区分性训练的实验结果,序贯区分性训练使基线系统词错率下降了 10.4%,总体实际查询项权重代价提升了 16.5%,具有良好的效果。同时,对于 LSTM+SP 声学模型系统,序贯区分性训练提升效果甚微,而 LSTM+M+A+D 声学模型系统则具有很好的性能。

表 6 基于 sMBR 准则的序贯区分性训练模型系统性能

声学模型	词错率/%	实际查询项权重代价		
		集内词	集外词	所有查询词
LSTM+D	54.9	0.394 2	0.157 4	0.371 2
LSTM+A+D	57.0	0.375 0	0.119 5	0.351 2
LSTM+M+D	46.1	0.487 6	0.231 7	0.463 7
LSTM+M+A+D	**44.8**	**0.490 7**	**0.228 4**	**0.466 2**

注:"D"表示声学建模使用了半监督训练基于 sMBR 准则的序贯区分性训练方法。

从现有 ASR 系统中挑选若干进行系统融合,选取目前性能最好的独立系统 LSTM+M+A+D 声学模型系统以及与之具有结构差异且性能较好的 SGMM、LSTM、LSTM+M、LSTM+M+A 和 LSTM+M+S+A 共 6 个声学模型系统进行融合,各系统权重均设置为 1/6。系统融合性能如表 7 所示。由表 7 结果可知,连续语音识别性能和关键词检索性能均优于之前所有的单独系统,相比基线系统连续语音识别词错率下降了 29.9%,关键词检索总体实际查询项权重代价提升了 60.3%。

表 7 系统融合性能

声学模型	词错率/%	实际查询项权重代价		
		集内词	集外词	所有查询词
6 个系统融合	43.0	0.491 5	0.028 4	0.517 7

4 结 语

本文提出一种采用长短时记忆网络的低资源语音识别方法,该方法是 LSTM 声学建模方法在低资源下的扩展。实验表明,利用其他语言数据提取的多语言瓶颈特征具有更好的表征力。半监督训练和基于语速扰动的数据扩展也有效提升了系统的识别性能。LSTM 在前期训练完成后再经过序贯区分性训练能得到更准确的识别结果。最后系统融合可以充分利用不同系统间的差异性,得到最佳识别效果。

参考文献:

[1] 刘加,张卫强. 低资源语音识别若干关键技术研究进展[J]. 数据采集与处理,2017,32(2):205-220.
LIU Jia, ZHANG Weiqiang. Research progress on key technology of low resource speech recognition [J]. Data Acquisition & Processing, 2017, 32(2): 205-220.

[2] IARPA. The babel program [EB/OL]. [2017-01-15]. https://www.iarpa.gov/index.php/research-programs/babel.

[3] CAI M, LV Z, SONG B, et al. The THUEE system for the openKWS14 keyword search evaluation [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2015: 4734-4738.

[4] DO V H, XIAO X, XU H, et al. Multilingual exemplar-based acoustic model for the NIST Open KWS 2015 evaluation [C]// Asia-Pacific Signal and Information Processing Association Summit and Conference. Piscataway, NJ, USA: IEEE, 2015: 594-598.

[5] CHEN I F, NI C, LIM B P, et al. A keyword-aware language modeling approach to spoken keyword search [J]. Journal of Signal Processing Systems, 2016, 82(2): 197-206.

[6] 袁胜龙. 资源受限情况下基于 ASR 的关键词检索研究[D]. 合肥:中国科学技术大学,2016:17-18.

[7] 刘迪源. 基于 BN 特征的声学建模研究及其在关键词检索中的应用[D]. 合肥:中国科学技术大学,2015:12-14.

[8] NIST. DRAFT KWS16 keyword search evaluation plan [EB/OL]. [2017-01-15]. https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf.

[9] 俞栋,邓力. 解析深度学习:语音识别实践[M]. 北京:工业出版社,2016:78-84.

[10] KNILL K M, GALES M J F, RATH S P, et al. In-

vestigation of multilingual deep neural networks for spoken term detection [C]// Automatic Speech Recognition and Understanding. Piscataway, NJ, USA: IEEE, 2013: 138-143.

[11] VESELY K, HANNEMANN M, BURGET L. Semi-supervised training of deep neural networks [C] // 2013 Automatic Speech Recognition and Understanding. Washington, DC, USA: IEEE Computer Society, 2013: 267-272.

[12] KO T, PEDDINTI V, POVEY D, et al. Audio augmentation for speech recognition [C]// Proceedings of the Annual Conference of the International Speech Communication Association. Lous Tourils, Baixas, France: International Speech and Communication Association, 2015: 3586-3589.

[13] VESELY K, GHOSHAL A, BURGET L, et al. Sequence-discriminative training of deep neural networks [C]// Proceedings of the Annual Conference of the International Speech Communication Association. Lous Tourils, Baixas, France: International Speech and Communication Association. 2013: 2345-2349.

[14] KINGSBURY B. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling [C]// Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE, 2009: 3761-3764.

[15] KINGSBURY B, SAINATH T N, SOLTAU H. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization [C] // Proceedings of the 13th Annual Conference of the International Speech Communication Association 2012. Lous Tourils, Baixas, France: International Speech and Communication Association, 2012: 10-13.

[16] MAMOU J, CUI J, CUI X, et al. System combination and score normalization for spoken term detection [C] // Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE, 2013: 8272-8276.

[17] GOEL V, KUMAR S, BYRNE W. Segmental minimum Bayes-risk decoding for automatic speech recognition [J]. IEEE transactions on Speech and Audio Processing, 2004, 12(3): 234-249.

[18] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit [EB/OL]. [2016-12-12]. http: // homepages. inf. ed. ac. uk/ag hoshal/pubs/asru11-kaldi. pdf.

[19] 陆梨花，张连海，陈琦. 基于加权有限状态转换器的语音查询项检索技术 [J]. 数据采集与处理，2015，30(2)：390-398.
LU Lihua, ZHANG Lianhai, CHEN Qi. Spoken term detection techniques based on weight finite-state transducer [J]. Data Acquisition & Processing, 2015, 30(2): 390-398.

[20] GHAHREMANI P, BABAALI B, POVEY D, et al. A pitch extraction algorithm tuned for automatic speech recognition [C]// Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE, 2014: 2494-2498.

[21] 吴蔚澜，蔡猛，田垚，等. 低数据资源条件下基于 Bottleneck 特征与 SGMM 模型的语音识别系统 [J]. 中国科学院大学学报，2015，32(1)：97-102.
WU Weilan, CAI Meng, TIAN Yao, et al. Bottleneck features and subspace Gaussian mixture models for low-resource speech recognition [J]. Journal of University of Chinese Academy of Sciences, 2015, 32(1): 97-102.

[本刊相关文献链接]

宋青松，田正鑫，孙文磊，等. 用于孤立数字语音识别的一种组合降维方法. 2016，50(6)：42-46. [doi：10. 7652/xjtuxb201606007]

范正光，屈丹，闫红刚，等. 借助音频数据的发音字典新词学习方法. 2016，50(6)：75-82. [doi：10. 7652/xjtuxb201606012]

陈斌，胡平舸，屈丹. 子空间域相关特征变换与融合的语音识别方法. 2016，50(4)：60-67. [doi：10. 7652/xjtuxb201604010]

刘晓明，班超帆，冯晓荣. 失真控制下的短时谱估计语音增强算法. 2011，45(8)：78-84. [doi：10. 7652/xjtuxb201108014]

张琦，王霞，孙焘. 高斯混合模型下的残留回波抑制算法. 2013，47(4)：11-16. [doi：10. 7652/xjtuxb201304003]

曾立飞，刘观伟，毛靖儒，等. 调节阀内流型分布及利用声音突变判别流型转变的方法. 2015，49(5)：116-121. [doi：10. 7652/xjtuxb201505018]

张选平，王旭，邵利平，等. 音频数据加密的二值音频可听密码方案. 2012，46(8)：27-32. [doi：10. 7652/xjtuxb201208005]

郭一鸣，彭华，张冬玲，等. 联合串行干扰抵消与因子图的单通道混合信号盲分离算法. 2015，49(10)：130-135. [doi：10. 7652/xjtuxb201510021]

张东伟，郭英，齐子森，等. 采用空间极化时频分布的跳频信号多参数联合估计算法. 2015，49(8)：17-23. [doi：10. 7652/xjtuxb201508004]

郝雯洁，齐春. 一种鲁棒的稀疏信号重构算法. 2015，49(4)：98-103. [doi：10. 7652/xjtuxb201504016]

（编辑　刘杨）