ISSN 1000-0054 清华大学学报（自然科学版） 2018 年 第 58 卷 第 4 期 2/17
CN 11-2223/N J Tsinghua Univ (Sci & Technol), 2018, Vol. 58, No. 4 342-346

跨语言声学模型在维吾尔语语音识别中的应用

努尔麦麦提·尤鲁瓦斯[1]， 刘俊华[2]， 吾守尔·斯拉木[1]，
热依曼·吐尔逊[1]， 达吾勒·阿布都哈依尔[1]

（1. 新疆大学 信息科学与工程学院，乌鲁木齐 830046；2. 科大讯飞股份有限公司，合肥 230088）

摘 要：对维吾尔语而言，由于数据采集和标注存在各种困难，用于训练声学模型的语音数据不够充分。为此，该文研究了基于长短期记忆网络的跨语言声学模型建模方法，利用汉语庞大的训练数据训练深度神经网络声学模型，然后将网络的输出层权重去掉，用随机化的方式产生与维吾尔语输出层对应的权重值，采用反向传播的方式，利用维吾尔语语音数据更新所有权重来训练维吾尔语声学模型。实验结果表明：该方法使维吾尔语转写和听写识别错误率分别比基线系统相对降低了 20%和 30%。该方法利用汉语大数据来训练神经网络的隐藏层，使维吾尔语声学模型能在一个较好的初始权重网络上进行训练，增强了网络的鲁棒性。

关键词：声学模型；维吾尔语；跨语言；长短期记忆

中图分类号：TP391.4 **文献标志码**：A

文章编号：1000-0054(2018)04-0342-05

DOI：10.16511/j.cnki.qhdxxb.2018.22.020

Crosslingual acoustic modeling in Uyghur speech recognition

NURMEMET Yolwas[1], LIU Junhua[2], WUSHOUR Silamu[1],
REYIMAN Tursun[1], DAWEL Abilhayer[1]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
2. iFLYTEK Co., Ltd., Hefei 230088, China)

Abstract: The Uyghur language has a little speech data for training acoustic models due to various data acquisition and annotation difficulties. This paper describes a modeling method for crosslingual acoustic models based on long short-term memory models. Mass Chinese language training data is used to train a deep neural network acoustic model. The network output layer weights are then randomly modified to create the output layer for the Uyghur language. A Uyghur language acoustic model is then trained using Uyghur language speech data to update all the weights. Tests show that this method reduces the word error rates of the Uyghur language transcription and dictation recognition by 20% and 30% than the baseline system. Thus, this method improves the Uyghur language acoustic model with better initial weights from the Chinese language data to train hidden layers in the neural network, and enhances the network robustness.

Key words: acoustic model; Uyghur; crosslingual; long short-term memory

最近在语音识别领域，含有多个隐藏层的前向神经网络即深度神经网络（deep neural network，DNN）击败了传统的 Gauss 混合模型（Gaussian mixture model，GMM），显著提升了汉语、英语等的语音识别率；其后，深度循环神经网络（recurrent neural network，RNN）被应用于语音识别任务，其语音识别错误率明显低于 DNN。尽管深度神经网络在维吾尔语大词汇量连续语音识别中的应用取得了一些进展[1-2]，但是训练深度神经网络各层的参数需要大规模有标注的语音数据。对于很多小语种而言，由于数据采集和标注存在各种困难，用于训练的语音数据往往没有英语、汉语那么多，因此如何利用英语、汉语等的大规模训练数据来提升小语种声学模型的识别性能成为了重要的研究课题。

跨语言声学模型建模的主要思想是将在资源丰富的语言数据上得到的知识移植到资源缺乏的语言上。目前，跨语言声学建模已取得很多成果。在汉语方言普通话语音识别的声学建模问题上，研究者们通过少量方言普通话数据，根据标准普通话和方言普通话的发音差异，利用距离度量作为生成准则建立声学模型，提高了方言普通话的识别率[3]。文[4]使用多种语言语音数据来建立包含各种语言音素集的通用声学模型，然后将通用声学模型自适应

收稿日期：2017-09-30
基金项目：国家自然科学基金项目（61363063，U1603262，61462084）；新疆维吾尔自治区重点实验室项目（2015KL013）
作者简介：努尔麦麦提·尤鲁瓦斯（1980—），男，副教授。
E-mail：nurmemet@xju.edu.cn

到新语言上。此方法的缺点是创建上下文相关模型时参数无法充分训练，导致识别率下降。然而，基于结构性转移学习的子空间 Gauss 混合模型(subspace Gaussian mixture model，SGMM)[5]和深度神经网络方法在文[4]的基础上证明了声学模型可以实现跨语言迁移学习。SGMM 跨语言迁移学习方法的主要思想是将 GMM 的参数分为状态相关的参数和状态无关的全局参数。全局参数不依赖于任何语言音素或隐 Markov 模型(hidden Markov model，HMM)中的状态，这些全局参数可以被不同语言共享。在新语言只有少量语音数据的情况下，只需训练状态相关的参数，然后通过共享全局参数来估计新语言的声学模型[6]。对于深度神经网络的迁移学习，研究者们提出了将资源丰富的语言的语音数据输入到神经网络中，将神经网络的输出层或者隐藏层的输出进行处理获取 Tandem 或 Bottleneck 特征，然后用获取的 Tandem 和 Bottleneck 特征与目标语言的 GMM 或 SGMM 相结合的方法[7-8]；还有研究者提出了使用资源丰富的语言数据来预训练深度神经网络的隐藏层，再用少量目标语言数据训练网络的方法[9]。由于人类语言发音共享相似的声学空间，语言之间存在很多相似的声学单元，因此研究者们提出了通过设计音素映射层[10]或状态映射层[11]，用已充分训练的源语言声学模型来创建目标语言声学模型的方法。最近，有研究者提出了在深度神经网络中设置多个 Softmax 输出层的方法，每个输出层对应一种语言，这些输出层共享深度神经网络的隐藏层参数，从而提高对目标语言的识别率[12]。

针对维吾尔语标注困难、声学模型语音训练数据有限的问题，本文利用跨语言声学模型建模思想，首先设计双向长短期记忆(long-short term memory，LSTM)网络和 RNN 结构，利用汉语庞大的训练数据训练深度神经网络声学模型，将获得的声学模型和相应参数作为维吾尔语声学模型的初始模型。然后，根据维吾尔语自身的特点，通过修改 LSTM 输出层的结构，利用维吾尔语标注数据重新训练来获得维吾尔语语音识别的声学模型。最后，在维吾尔语转写任务和听写任务上对所建立的维吾尔语声学模型进行测试实验。

1 维吾尔语的音位系统

现代维吾尔标准语的音位系统里总共有 32 个音位。其中：8 个为元音音位，24 个为辅音音位。在维吾尔语里，音素的长度主要表现在元音上，特别是低元音上，而高元音几乎不存在长短方面的区别。维吾尔语的音系现象包括元音和谐、元音清化、语音弱化、辅音同化、语音脱落等。在音节结构上，维吾尔语也有自己的特点。维吾尔语的固有音节结构里存在 6 种音节类型，即 V、VC、CV、CVC、VCC、CVCC。其中：字母 V 表示元音，C 表示辅音。除此之外，还存在 CCV、CCVC、CCVCC、CVV、CVVC 这 5 种从外语借词后在现代维吾尔语里出现的音节结构类型。

2 基于深度循环神经网络的声学模型

RNN 与前馈深度神经网络都是人工神经网络，在结构上它们的主要区别在于 RNN 的同一隐藏层节点间有一个反馈的循环连接。由 DNN 的特性可以知道，作为前馈神经网络，DNN 只允许固定窗长度输入，只能将当前输入映射到输出；而 RNN 通过增加一个循环连接，可以将历史输入记录下来，保存在网络的中间状态中，并用于影响网络的输出。也就是说，RNN 原则上是将当前输入和历史输入一起映射到每一个输出。由此可知，DNN 的记忆受限于输入窗长度，而 RNN 的记忆则是不受限制的。

尽管 RNN 很早就在音素识别[13]中取得了成功，但由于其训练的复杂性以及存在梯度消失问题，很难应用在大规模的语音识别任务上。为此，研究者们提出长短时记忆模型来解决传统简单 RNN 存在的梯度消失等问题，使得 RNN 框架可以应用到语音识别领域并取得了超越 DNN 的效果。但 LSTM 无法实现借助未来信息进行辅助判决，为此研究者们提出了双向 LSTM。双向 LSTM 利用前后双向的 LSTM 结构，不仅能够利用历史信息，还可以利用未来信息进行辅助判决，因此它的模型描述能力比单向 LSTM 更强。采用双向 LSTM 的双向 RNN 声学模型结构如图 1 所示。

由图 1 可知，本文采用双向 RNN 结构对 HMM 中三音素的绑定状态进行建模。RNN 输出层对多层的非线性特征变换进行 Softmax 处理。隐藏层 $\overrightarrow{h}$ 和 $\overleftarrow{h}$ 分别表示双向 RNN 中前向层和后向层，并采用如图 2 所示的 LSTM 结构。

LSTM 隐藏层由输入门、输出门和遗忘门控制信息的读入、写出和重置。当遗忘门的值为 1 时，

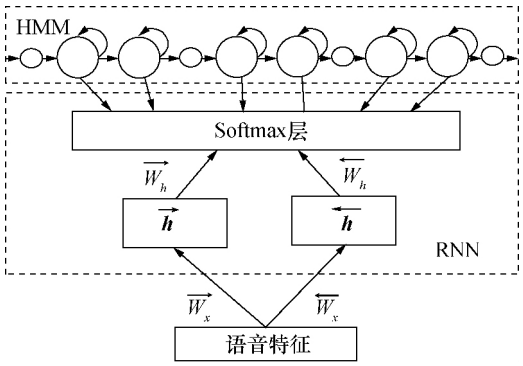

图 1 双向 RNN 声学模型示意图

可实现历史信息的无损记忆。LSTM 中各门使用 Sigmoid 激活函数，而输入门和记忆单元通常会使用 tanh 函数来转换。LSTM 的记忆单元和各门可以用下列等式来定义：

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{X}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{C}_{t-1} + \boldsymbol{b}_i), \quad (1)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{X}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{C}_{t-1} + \boldsymbol{b}_f), \quad (2)$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t\boldsymbol{C}_{t-1} + \boldsymbol{i}_t\tanh(\boldsymbol{W}_{xc}\boldsymbol{X}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c), \quad (3)$$

$$\boldsymbol{o}_t = \sigma(W_{xo}\boldsymbol{X}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{C}_t + \boldsymbol{b}_o), \quad (4)$$

$$\boldsymbol{h}_t = \boldsymbol{O}_t\tanh(\boldsymbol{C}_t). \quad (5)$$

输入门 i、遗忘门 f、输出门 o 这 3 个门的输入都是 t 时刻输入向量 $\boldsymbol{X}_t$。在 $t-1$ 时刻，隐藏层向量为 $\boldsymbol{h}_{t-1}$，记忆单元状态为 $\boldsymbol{C}_{t-1}$。W 为网络参数矩阵，b 为偏置向量，tanh( )为转换函数，σ( )为 Sigmoid 激活函数。

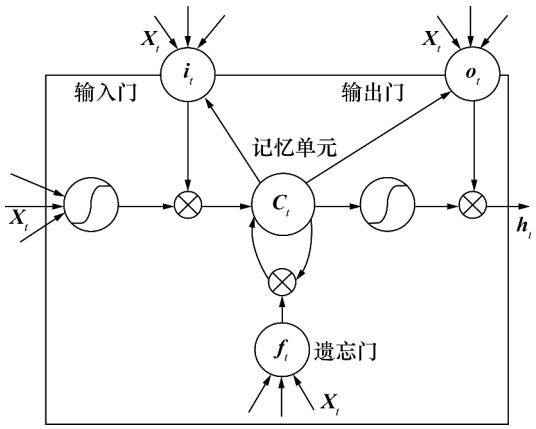

图 2 LSTM 结构示意图(包含 1 个记忆单元)

3 跨语言声学模型训练方法

本文采用的跨语言声学模型训练方法的具体做法是：利用汉语庞大的训练数据训练神经网络声学模型，然后将神经网络的输出层权重去掉，用随机化的方式产生与维吾尔语输出层对应的权重值，再采用反向传播算法，利用维吾尔语语音数据进行重新训练。该方法的优点在于充分利用了汉语大数据来训练神经网络的隐藏层，使模型具有较好的初始权重，然后让维吾尔语声学模型能够在一个具有较好的初始权重的神经网络上进行训练，从而增强网络的鲁棒性。

4 实验结果与分析

4.1 数据集

本文使用两种数据集分别进行维吾尔语语音转写任务和语音听写任务。两种数据集分别是基于电话信道的口语语音数据集和来自手机输入法的语音数据集。电话信道数据集包含 1 个训练集和 4 个测试集，训练集总共 425 h，测试集总共 8 h。每个测试集有 2 h 有效语音，来自不同时间、不同群体，说话风格存在差异。电话信道数据主要取自新疆地区，以 wav 格式保存，所有数据通过人工进行了标注。来自手机语音输入法的语音数据集包含 1 个 600 h 的训练集和 2 个测试集，测试集共 4 h。训练集和测试集都来自手机语音输入法数据，对所有数据通过人工进行了标注。

为了验证跨语言声学模型建模方法在维吾尔语语音识别中的应用效果，本文使用了汉语语音数据集。该数据集也分别来自电话信道和语音输入法，分别包括 6 000 h 和 8 000 h 的训练数据。

4.2 声学模型配置

将本文方法与 4 种常用的声学模型的性能进行对比：1) GMM-HMM 声学模型在建立时对训练语音数据提取 39 维 mel 频率倒谱系数(mel-frequency cepstral coefficient，MFCC)特征，基线声学模型将维吾尔语音素作为基元，使用最大似然估计(maximum likelihood estimation，MLE)准则训练，然后将训练出来的单音素模型扩展成上下文相关的三音素模型。2) DNN-HMM 模型的输入为 24 维 FBank 特征加上一阶差分和二阶差分，将其前后各取 9 帧组成 648 维输入节点。此模型共有 6 个隐藏层，每层有 2 048 个节点。输出层对应的聚类后的状态标签有 9 000 个节点。3) LSTM-HMM 声学模型中单向 LSTM 网络包含 1 个输入层，输入层的节点对应 40 维 FBank 特征，扩了 5 帧，共 200 个节点；3 个隐藏层，每层有 2 048 个节点。4) BLSTM-HMM 和 Crosslingual 声学模型采用的 BLSTM 网络结构为 3 层双向 LSTM，每层前向有 1 024 个节点、后向有 1 024 个节点，每个节点包含

一个记忆单元。Sigmoid 函数作为隐藏层的激活函数，输出层分类用 Softmax 函数，其他配置与单向 LSTM 和 DNN-HMM 类似。BLSTM-HMM 和 Crosslingual 声学模型的区别在于 Crosslingual 先采用汉语语音数据集进行网络训练，然后利用维吾尔语语音数据进行重新训练，而 BLSTM-HMM 只利用维吾尔语语音数据进行训练。

DNN-HMM、LSTM-HMM、BLSTM-HMM、Crosslingual 声学模型采用 minibatch 随机梯度下降(stochastic gradient descent，SGD)算法进行训练，选用交叉熵和最小音素错误率作为目标函数。由于数据量比较充足，因此训练 Crosslingual 声学模型时没有进行预训练，每次迭代对整个网络的权重进行更新。解码时采用三元语法模型，使用基于加权有限状态转换器(weighted finite state transducer，WFST)的静态解码框架。

4.3 实验结果

本文对不同声学模型利用基于电话信道数据集的 4 个测试集(T1、T2、T3、T4)进行了维吾尔语语音转写任务的性能对比，各模型的词识别错误率(WER)如表 1 所示。转写不需要实时完成，而听写是有实时性要求的，一般 BLSTM 不满足实时性要求，因此在维吾尔语语音听写任务中没有考察 BLSTM-HMM。本文对不同声学模型利用语音输入法的两个测试集(T5、T6)进行了维吾尔语语音听写任务的性能对比，结果如表 2 所示。

由表 1 和 2 可以看出，基于深度神经网络的声学建模方法的性能优于基于 GMM-HMM 的，从词识别错误率来看，Crosslingual 模型比 GMM-HMM 在转写和听写任务中 WER 分别下降了 20% 和 30%，这说明在大规模数据上基于深度神经网络的声学模型比 GMM-HMM 模型在性能上有很大提高。基于深度神经网络的声学模型中最好的 LSTM-HMM 比 DNN-HMM 的 WER 在转写和听写任务中分别下降了 10% 和 13.8%，这说明 LSTM 对上下文建模能力更强，再次证明了 LSTM 声学模型在语音识别中具有较好的效果。采用汉语语音数据训练出来的 Crosslingual 声学模型比利用维吾尔语语音数据训练出来的 LSTM-HMM 性能好，在转写和听写任务中词识别错误率分别下降了 4%和 6%，这说明可以利用资源丰富的语言的语音数据来提高维吾尔语声学模型的性能。但是，该性能提高不是很大，这可能是因为维吾尔语语音数据量已经有几百小时了，数据量并不算很少，所以 Crosslingual 比 LSTM-HMM 的性能提升不是很显著。另一个可能的原因是维吾尔语和汉语属于不同语系，两种语言中发音相似的音素较少，这也限制了声学模型性能的进一步提高。

从整体实验结果来看，各模型在语音转写任务上的性能比听写任务要差，这是由于语音转写任务数据来自电话信道，语音质量比较差，语音内容中说话风格随意，并且口语化严重，因此导致识别错误率较高。

表 1 各种声学模型在语音转写任务中的性能

声学模型	WER/%				
	T1	T2	T3	T4	平均
GMM-HMM	41.04	41.68	49.44	48.83	45.25
DNN-HMM	37.13	38.84	45.79	46.31	42.02
LSTM-HMM	32.67	36.16	43.04	43.54	38.86
BLSTM-HMM	31.55	35.56	41.38	42.15	37.66
Crosslingual	30.40	34.04	39.12	40.46	36.01

表 2 各种声学模型在语音听写任务中的性能

声学模型	WER/%		
	T5	T6	平均
GMM-HMM	27.33	28.75	28.04
DNN-HMM	23.52	24.39	23.96
LSTM-HMM	20.53	20.74	20.64
Crosslingual	18.99	19.65	19.32

5 结束语

本文针对维吾尔语语音数据不足的问题，研究了跨语言声学模型在维吾尔语语音识别中的应用，采用了基于长短期记忆网络的跨语言声学模型建模方法和跨语言声学模型训练方法，建立了 GMM-HMM、DNN-HMM、LSTM-HMM、BLSTM-DNN、Crosslingual 等声学模型，并对各声学模型在维吾尔语语音转写和语音听写测试任务上的识别性能进行了分析。从本文实验结果可以看出，跨语言声学模型建模方法提升了维吾尔语声学模型的性能，使用其他大语种大规模语料库数据训练声学模型能够获得比较稳定的发音描述模型，在此基础上基于维吾尔语训练数据进行自适应训练可以使该模型在维吾尔语上具有更好的区分能力。本文作者认为将汉语和维吾尔语的声学知识与跨语言声学模型的建模方法相结合可以进一步降低语音识别错误率。此外，

本文方法不但对维吾尔语有效，而且对语音资源相对缺乏的哈萨克语、柯尔克孜语等的语音识别研究具有重要参考意义。

参考文献 (References)

[1] **麦麦提艾力·吐尔逊，戴礼荣. 深度神经网络在维吾尔语大词汇量连续语音识别中的应用**[J]. **数据采集与处理**，2015，30(2)：365-371.
MAIMAITIAILI T，DAI L R. Deep neural network based Uyghur large vocabulary continuous speech recognition [J]. Journal of Data Acquisition and Processing，2015，30(2)：365-371. (in Chinese)

[2] **其米克·巴特西，黄浩，王羡慧. 基于深度神经网络的维吾尔语语音识别**[J]. **计算机工程与设计**，2015，36(8)：2239-2244.
QIMIKE B，HUANG H，WANG X H. Uyghur speech recognition based on deep neural network [J]. Computer Engineering and Design，2015，36(8)：2239-2244. (in Chinese)

[3] **刘林泉，郑方，吴文虎. 基于小数据量的方言普通话语音识别声学建模**[J]. **清华大学学报(自然科学版)**，2008，48(4)：604-607.
LIU L Q，ZHENG F，WU W H. Small dataset-based acoustic modeling for dialectal Chinese speech recognition [J]. Journal of Tsinghua University (Science and Technology)，2008，48(4)：604-607. (in Chinese)

[4] SCHULTZ T，WAIBEL A. Experiments on cross-language acoustic modeling [C]// The 7th European Conference on Speech Communication and Technology. Aalborg，Denmark，2001：2721-2724.

[5] POVEY D，BURGET L，AGARWAL M，et al. The subspace Gaussian mixture model：A structured model for speech recognition [J]. Computer Speech & Language，2011，25(2)：404-439.

[6] BURGET L，SCHWARZ P，AGARWAL M，et al. Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models [C] // IEEE International Conference on Acoustics Speech and Signal Processing. Dallas，USA，2010：4334-4337.

[7] STOLCKE A，GREZL F，HWANG M Y，et al. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptron [C]// IEEE International Conference on Acoustics，Speech and Signal Processing. Toulouse，France，2006：321-324.

[8] VESELÝ K，KARAFIÁT M，GRÉZL F，et al. The language-independent bottleneck features [C]// 2012 Workshop on Spoken Language Technology. Miami，USA，2012：336-341.

[9] SWIETOJANSKI P，GHOSHAL A，RENALS S. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR [C]// 2012 Workshop on Spoken Language Technology. Miami，USA，2012：246-251.

[10] SIM K C，LI H. Context-sensitive probabilistic phone mapping model for cross-lingual speech recognition [C]// 9th Annual Conference of the International Speech Communication Association. Brisbane，Australia，2008：2715-2718.

[11] DO V H，XIAO X，CHNG E S，et al. Context dependant phone mapping for cross-lingual acoustic modeling [C]// 2012 8th International Symposium on Chinese Spoken Language Processing. Hong Kong，China，2012：16-20.

[12] HUANG J T，LI J，YU D，et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers [C]// IEEE International Conference on Acoustics，Speech and Signal Processing. Vancouver，Canada，2013：7304-7308.

[13] ROBINSON A J. An application of recurrent nets to phone probability estimation [J]. IEEE Transactions on Neural Networks，1994，5(2)：298-305.

(责任编辑 李丽)