基于先验概率线性插值的声学模型自适应方法 *

王丽[1]，张震[2]，张鹏远[1,3]，颜永红[1,3,4]
(1.中国科学院声学研究所 语言声学与内容理解重点实验室，北京 100190；2.国家计算机网络应急技术处理协调中心，北京 100029；3.中国科学院新疆理化技术研究所 新疆民族语音语言信息处理实验室，乌鲁木齐 830011；4.中国科学院大学，北京 100049)

文 摘：基于 DNN 的声学模型自适应算法是语音识别领域中重要的研究方向，自适应技术是通过少量的自适应数据优化基线模型，解决由于目标测试数据与训练数据不匹配而导致的识别性能下降的问题。实际上，少量的自适应数据几乎无法描述所有目标测试数据的分布情况，因此通过自适应数据统计得到的先验概率同样缺乏通用性。文章中针对这个问题，提出了基于先验概率线性插值的声学模型自适应方法，该方法将自适应数据的先验与基线模型的先验进行线性插值，使先验知识不仅融合了测试数据信息而且更具有通用性。实验证明，通过线性插值的先验可以有效改善识别性能。

关键词：声学模型自适应；深度神经网络；先验概率
分类号 1；分类号 2

自 2010 年微软研究人员提出采用上下文相关的深度神经网络（Deep Neural Network, DNN）建模技术[1]以来，由于其显著的性能优势，DNN 引起了语音识别领域研究人员的广泛关注。目前基于 DNN-HMM 的语音识别系统虽然可以取得比较好的识别性能，但是当测试数据与训练数据由于说话人、信道、环境等原因导致不匹配时，识别性能会大幅下降。因此如何提高声学模型的鲁棒性和泛化能力，使声学模型可以适应不同的测试数据，仍然是语音识别领域中的难点。声学模型的自适应技术是很多语音识别人员非常关注的研究方向。基于 DNN-HMM 的自适应策略有重训练[2]、在神经网络中添加线性变换层[3-4]以及基于正则化的自适应算法[5-7]等，另外大量实验证明，说话人自适应技术，i-Vector[8]及 Speaker Code[9]可以有效的改善针对特定说话人的识别性能。

目前声学模型自适应方法主要是利用少量与测试环境相匹配的自适数据优化基线声学模型，从而达到提高语音识别性能的目的。实际上，少量的自适应数据几乎无法描述所有目标测试数据的分布情况，因此通过自适应数据统计得到的先验概率同样缺乏通用性。针对数据稀疏引起的状态先验分布不均的问题，我们探究了先验概率的线性插值计算方法，将自适应数据的先验与基线模型的先验进行线性插值，更新后的先验不仅融合了自适应数据的分布信息，而且更具有通用性。

本文首先介绍了实验中用到的网络结构及训练准则；然后探究了声学模型常用的自适应算法；在此基础上研究了先验概率的计算方法以及先验概率的线性插值计算方法；然后通过实验证明，通过线性插值的先验可以有效改善识别性能；最后是结论。

1 语音识别系统

目前基于深度神经网络的语音识别系统因为具有显著的识别性能优势，而被广泛的应用。根据不同的网络结构，神经网络可以分为①前馈神经网络（Feedforward Neural Network, FNN）、②卷积神经网络（Convolutional Neural Networks, CNN）和③递归神经网络（Recurrent Neural Networks, RNN）等。LSTM（Long Short-Term Memory）[10]是一种特殊的 RNN 网络，可以学习长时间的依赖关系，因此可以很好的处理序列化数据，是当前常用的网络结构。

1.1 LSTM 的网络结构

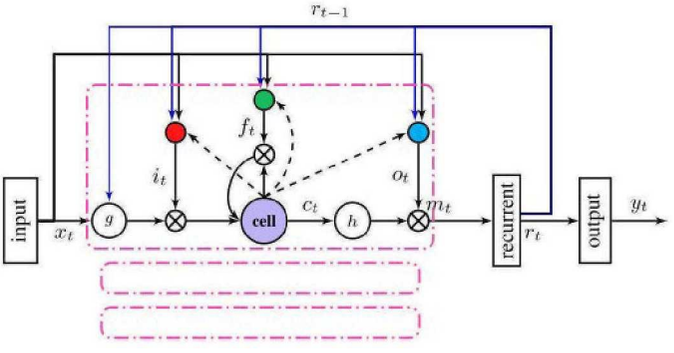

图 1 LSTM 结构（粉色区域指 LSTM 记忆单元）

LSTM 的网络结构包含输入层、隐藏层、输出层。输入层输入语音的特征观察值；隐藏层一般包含多个非线性层，每个非线性层都是由若干

* 基金项目：本论文工作得到国家自然科学基金（11590770-4,U1536117,11504406,11461141004），国家重点研发计划重点专项（2016YFB0801203,2016YFB0801200）和新疆维吾尔自治区科技重大专项（2016A03007-1）经费资助
作者简介：王丽（1985.11），女（汉），山东省临沂市沂水县，助理研究员；E-mail：wangli@hccl.ioa.ac.cn

图 1 所示的记忆单元组成；输出层的节点对应着上下文相关的 triphone 的三状态。

假定输入的特征序列为 $x=(x_1,.....,x_T)$，输出序列为 $y=(y_1,.....,y_T)$，则 LSTM 的前向计算公式为：

$$i_t=\sigma\left(W_{ix}x_t+W_{im}m_{t-1}+W_{ic}c_{t-1}+b_i\right) \quad (1)$$

$$f_t=\sigma\left(W_{fx}x_t+W_{fm}m_{t-1}+W_{fc}c_{t-1}+b_f\right) \quad (2)$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g\left(W_{cx}x_t+W_{cm}m_{t-1}+b_c\right) \quad (3)$$

$$o_t=\sigma\left(W_{cx}x_t+W_{om}m_{t-1}+W_{oc}c_t+b_o\right) \quad (4)$$

$$m_t=o_t\odot h\left(c_t\right) \quad (5)$$

$$y_t=\phi\left(W_{ym}m_t+b_y\right) \quad (6)$$

式中，W 为权重矩阵，b 为偏差向量，$\odot$ 为点乘操作，g,h 分别表示激活函数，本文中采用的是 tanh 函数，ϕ 为 softmax 激活函数。

为了充分利用未来的信息，本文采用了双向 LSTM 网络。

1.2 神经网络的训练准则

交叉熵训练准则的目标函数如式(7)所示，式中 N 表示训练样本的数目；$\hat{p}\left(y|X_t\right)$ 表示标注概率，取值为 0 或者 1，训练数据的标注是通过强制对齐技术得到；$p\left(y|X_t\right)$ 表示通过 DNN 计算得到的概率。

$$F(W)=\sum_{t=1}^{N}\sum_{y=1}^{S}\hat{p}\left(y|X_t\right)\log p\left(y|X_t\right) \quad (7)$$

网络的更新公式如式(8)所示，η 表示学习率。根据目标函数不断更新权重参数，寻找使目标函数达到最小时的权重参数。

$$W^l(t)=W^l\left(t-1\right)-\eta\frac{\partial F}{\partial W^l},1\le l\le L \quad (8)$$

2 基于 DNN 的自适应算法

虽然现在基于 DNN 的语音识别已经很完善，但是在实际应用中，当训练数据与测试数据不匹配时，通常情况下是由说话人、信道、麦克风及环境变量等引起，仍旧会导致识别性能大幅度下降。为了改善由于数据不匹配引起的性能衰减的问题，声学模型的自适应技术变得尤为重要，也得到了人们的广泛关注。

目前声学模型自适应技术分为：有监督自适应和无监督自适应。由于无监督自适应总会存在高于手工标注的识别错误，因此本文都是采用了有监督自适应技术。

2.1 神经网络的重训练自适应算法

神经网络的重训练自适应是一种最简单的自适应方法。它利用已经训练完成的基线声学模型和特定领域的训练数据，通过重新的训练过程，使声学模型与特定领域更加匹配。在重新训练过程中，训练准则与基线模型的训练准则是一致的，DNN的所有权重通过BP进行更新。重训练策略的缺点是容易过拟合。为了缓解重训练过程中的过拟合问题，目前有两种解决方法：第一种方法是被广泛采用的交叉验证方法，根据开发集上的验证结果决定何时停止模型的训练；第二种方法是融合基线声学模型的训练数据和自适应数据，这种方法的难点在于需要通过大量的实验筛选出合理的数据融合比例，如果基线训练数据太多，会淹没自适应数据的特性，如果基线数据太少则达不到理想的效果。

2.2 相对熵正则化自适应算法

重训练自适应算法其实在本质上已经改变了基线模型学习到的数据特征，尤其当自适应数据与基线训练数据风格及领域差距很大时。为了防止上述的问题，可对基线声学模型进行相对熵正则化自适应算法。相对熵正则化自适应算法利用相对熵衡量自适应数据与基线模型训练数据的差异性，并将这种差异信息加入到训练准则中。

正则化的优化准则为如式(9)所示：

$$\hat{F}(W)=(1-\rho)F(W)+\rho\sum_{t=1}^{N}\sum_{y=1}^{S}p^{SI}\left(y|X_t\right)\log p\left(y|X_t\right) \quad (9)$$

式中 $p^{SI}\left(y|X_t\right)$ 是基线模型前向计算得到的后验概率，ρ 为正则化权重。上式通过整理可得到如下形式：

$$\hat{F}(W)=\sum_{t=1}^{N}\sum_{y=1}^{S}\tilde{p}\left(y|X_t\right)\log p\left(y|X_t\right) \quad (10)$$

式中 $\tilde{p}\left(y|X_t\right)=(1-\rho)\hat{p}\left(y|X_t\right)+\rho p^{SI}\left(y|X_t\right)$。由上式可以看出，正则化自适应的新目标概率是基线模型概率与强制对齐结果概率（标注概率）的线性插值。正则化权重可以根据自适应数据集和学习率的大小，使用开始集调整。当 $\rho=0$ 时，就等效于之前讲的重训练自适应算法；当 $\rho=1$ 时，模型就变为了基线模型。

3 先验概率

3.1 先验概率的计算方法

在 DNN-HMM 的语音识别系统中，声学模型可以提供 t 时刻特征矢量 O_t 由状态 S_j 产生的似然值：

$$p(o_t \mid s_j) = \frac{p(s_j \mid o_t)}{p(s_j)} \tag{11}$$

式中 $p(s_j \mid o_t)$ 为声学模型 DNN 输出的后验概率，$p(s_j)$ 为状态 S_j 的先验概率。

先验概率不仅描述了训练数据状态的分布情况，而且由上式可以看出，先验概率可以对后验概率进行规整，防止后验概率过于发散。

由于训练数据的限制性，先验概率很难准确的计算得到，一般都是通过估计得到。目前有两种常用的估计方法。第一种是通过统计强制对齐结果中各个状态的出现次数与所有状态的出现次数的比例得到；第二种方法是通过计算所有声学特征在状态 S_j 的后验概率得到[11]：

$$p(s_j) = \frac{1}{N}\sum_{i=1}^{N} p(s_j \mid o_i) \tag{12}$$

式中 N 为特征帧数。大量的实验证明，通过第二种方法可以取得更好的 WER 性能。

3.1 先验概率的自适应算法

一般情况下，自适应数据量都很少，数据分布发散，无法准确的表示目标测试数据的分布情况。那么，利用自适应数据统计得到的先验概率实际上是不能通用于目标测试数据的。

针对数据稀疏引起的状态先验分布不均的问题，我们尝试了基于线性插值的先验计算方法。

通常情况下，自适应的基线模型都是通过大量数据训练得到，泛化性好，表达能力强，由此得到的先验概率能很好地描述状态的分布情况。另外一方面，由自适应数据统计得到的先验概率更偏向于目标测试数据的分布，包含了测试数据的某些先验知识。因此通过式(13)的线性插值公式，可以充分利用这两者的优势，使先验更准确、泛化的描述实际测试数据的分布情况：

$$\hat{p}(s_j) = (1-\lambda)p(s_j) + \lambda p_{adjust}(s_j) \tag{13}$$

4 实验

4.1 实验数据及配置

本文的实验是针对实际应用中的特定任务进行的。基线声学模型的训练数据为 4000 小时电话交谈语音数据，自适应训练数据为 8 小时数字串朗读数据，这批数据来源于手机 APP 应用密码验证语音，验证码都是四位数字串。测试数据是与自适应训练数据同源的 262 条数字串数据，共 1028 个数字。

实验采用的音素集是 179 个中文带调音素集，由声、韵母音素组成。HMM 采用上、下文扩展的词间 triphone 建模，每个 HMM 三个状态。通过决策树方法将上下文相关的状态聚类成 8000 个状态单元。

实验采用的特征是 13 维 PLP 特征加上一维基频特征以及一维基频置信度，共 15 维基维特征，然后进行 3 阶差分扩展为 60 维。

实验采用的声学模型结构是充分考虑过去、未来信息的 BLSTM。输入特征维度为 300 维，由当前帧及前、后各 2 帧串联得到；隐藏层共三层，每个隐藏层包含 1024 个记忆单元，project 层的维度为 256；输出层采用 softmax 激活函数，包含 8000 个状态节点。

实验采用了重训练自适应算法。相对熵正则化自适应算法与重训练自适应算法的唯一区别在于训练准则上，实际上，重训练自适应是正则权重为 0 的正则化自适应方法，而且重训练方法易于实现，因此本文采用了重训练。自适应训练工具为 KALDI 开放环境，GPU 型号为 Tesla K20。

4.1 实验结果

表 1 记录了不同模型在测试集上的字错误率信息。通过表 1 的对比，可以发现以下两个结论：1)重训练自适应的方法可以改善声学模型的鲁棒性，提高声学模型在目标测试集上的识别性能；2)通过先验概率的线性插值，可以提高识别性能。为了探究插值系数对识别性能的影响，表 1 中进行了四组对比实验，当插值系数分别为 0.5 和 0.7 时可以取得最佳的识别性能。

表 1 不同模型的字错误率

模型	字错误率(%)
base	2.7
retrain	2.1
retrain_inter(0.3)	1.6
retrain_inter(0.5)	1.4
retrain_inter(0.7)	1.4
retrain_inter(0.9)	1.5

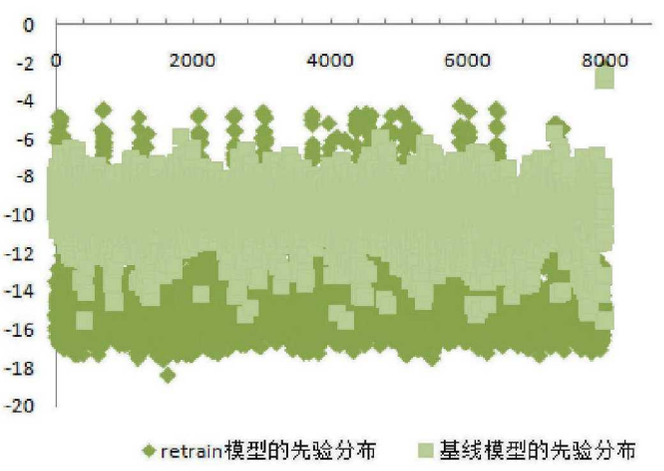

图 2　基线模型及 retrain 模型的先验分布图

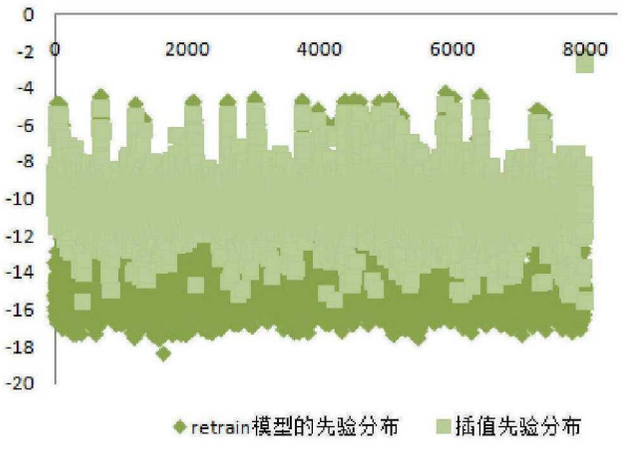

图 3　插值先验及 retrain 模型先验的分布图

图 2 描绘了通过不同数据统计得到的先验的分布图。浅颜色区域是基线模型的先验概率分布，深颜色是重训练模型的先验概率分布，可以看出，基线模型的先验分布更加紧凑，异常的状态比较少；重训练模型的先验分布广，发散状态比较多。这从另外一方面也可以说明通过少量数据统计得到的先验不具有代表性。图 3 是对比插值先验与 retrain 模型先验的分布，可以看出插值之后的先验分布有明显的收紧趋势。

4　结论

一般情况下，声学模型自适应的训练数据比较稀疏，通过自适应数据统计得到的先验概率分布无法准确的描述目标测试数据，而基线声学模型通常是由大量数据训练得到，其先验能够表示训练数据的分布信息。为了充分利用自适应数据的先验知识以及基线模型的通用性，尝试了基于先验概率线性插值的声学模型自适应方法，该方法将自适应数据的先验与基线模型的先验进行线性插值，使先验知识不仅融合了测试数据信息而且更具有通用性。上述实验证明，通过线性插值的先验可以有效改善识别性能。

在未来的工作中，会继续在不同的数据集上验证基于先验概率线性插值的声学模型自适应方法的通用性和有效性。

参 考 文 献

[1] Yu D, Seide F, Li G. Conversational speech transcription using context-dependent deep neural networks[C]. Edinburgh, Scotland: Omnipress. 2012:1-2.

[2] Yeming Xiao, Zhen Zhang, Shang Cai, et al. A initial attempt on task-specific adaptation for deep neural network-based large vocabulary continuous speech recognition[C]. Portland, Oregon: 2012.2574-2577.

[3] Gemello R, Mana F, Scanzio S, et al. Adaptation of Hybrid ANN/HMM Models Using Linear Hidden Transformations and Conservative Training[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. IEEE Xplore, 2006:I-I.

[4] Bo Li, Khe Chai Sim.. Comparison of Discriminative Input and Output Transformations for Speaker Adaptation in the Hybrid NN/HMM Systems[C]. Makuhari, Japan: 2010.526–529

[5] 张宇, 计哲, 万辛,等. 基于 DNN 的声学模型自适应实验研究[C]// 全国人机语音通讯学术会议. 2015.
Yu Zhang. An Experimental Study of Acoustics Model Adaptation Based on Deep Neural Networks[C]. Tijin:2015. (in china)

[6] Yu D, Yao K, Su H, et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:7893-7897.

[7] Tomashenko N, Khokhlov Y. Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing[C]// INTERSPEECH. 2014.

[8] Karanasou P, Wang Y, Gales M J F, et al. Adaptation of Deep Neural Network Acoustic Models Using Factorised I-Vectors[C]// INTERSPEECH. 2014.

[9] Xue S, Abdel-Hamid O, Jiang H, et al. Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(12):1713-1725.

[10] H. Sak, A. Senior, and F. Beaufays. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition[J]. Computer Science, 2014:338-342.

[11] V. Manohar, D. Povey, and S. Khudanpur. Semi-supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models[C]. INTERSPEECH 2015, pp. 2630–2634, 2015.

Adaptation of acoustics model based on linear interpolation of prior probability

Li Wang[1], Zhen Zhang[2], Pengyuan Zhang[1,3], Yonghong Yan[1,3,4]

1. Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

2. National Computer network Emergency Response technical Team/Coordination Center of China, Beijing 100029, China

3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of

Physics and Chemistry, Chinese Academy of Sciences, Urumchi 830011, China
4. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Adaptation of acoustic is very important in speech recognition. model aims at improving the performance of base acoustics model with small adaptation data set. In fact, small adaptive data can't represent distribution of test data., so prior probability based on adaptation data is incorrect used in test data. In this paper, we propose adaptation of acoustics model based on linear interpolation of prior probability. Based on linear interpolation, the prior probability combine information of adaptation data with train data of base model, thus, the prior can improve robustness of acoustic model. We experiment with special task and results show prior probability based on linear interpolation is helpful to speech recognition.

Key words: adaption of acoustics model; deep neural network; prior probability