基于跨语言特征的低资源维语声学模型训练

Low-resource Uyghur Acoustic Model Training based on Cross-lingual Features（CSLT-TRP-20150006）

殷实（Shi Yin）

2015/2/8

CSLT, RIIT, Tsinghua Univ.

1. 背景简介

随着中国语音技术的发展，汉语大词汇量连续语音识别已经达到了一个较好的识别性能，听写机、电话语音识别等也从实验室演示转移到实际应用的阶段。目前，语音识别市场由国外公司和机构占据了很大份额，系统以英语为主，我国紧跟语音识别领域的最新研究成果并基本与之保持同步。汉语语音技术的广泛应用，使大家看到语音技术的广大市场前景。

在新疆，维吾尔族是自治区的自治民族，少数民族尤其是维吾尔族在新疆人口中占有很大的比例。新疆地区官方语音是汉语和维语，由于维吾尔语语音特性，维、汉语之间的语言差异很大，正是这种少数民族的构成、人口与语言文字状况，使少数民族语言文字信息技术的开发与应用成为新疆信息化建设当中不可或缺的一个重要方面，也是国家信息化的基础之一。而研究维吾尔语的语音识别系统是新疆信息化建设的内容之一。具有重大的研究意义。到目前为止，国外无一机构（包括微软、IBM 等跨国公司）从事维语信息处理及维语语音识别系统的开发，所以目前国际上在此领域的研发也是一片空白。

同时，新疆地区的哈萨克民族和新疆周边的中亚国家，他们的语言文字和维吾尔语十分相似，维吾尔语的语音识别技术不仅在新疆有很广的应用前景，也能为这些语言的相关研究提供技术参考。语音识别系统广泛的应用市场和维吾尔族用户所占比例表明维吾尔文语音识别系统研究开发工作的必要性，其市场也是不容忽视的。

像维语这种小语种的语音识别是目前语音识别技术实用化必须要解决的一个难题。由于不同种类的语言音节本身就可能不一样，因此通常是通过针对不同的语言训练不同的声学模型来解决这种发音变化的问题。但是，在实际实现过程中，由于语音识别系统中的语音训练数据有限，不可能涵盖种类繁多的语言类型，从而会因为训练样本稀疏而导致声学模型与实际发音不匹配。

本研究关注维语语音识别系统中声学模型训练、建立过程。根据语音声学模型训练中的一些共通性，提出了基于跨语言的低资源维语声学模型训练方法。基本思想是以通用语言（例如汉语、英语等）的声学模型为基础，通过加入少量维语语料，重训练以改变声学模型中的参数，得到维语声学模型。这一方面利用了不同语言间的共享特征（如基本音素）等，实现少量语音条件下的快速训练。

同时，我们对比研究了直接使用少量数据训练的维语声学模型和利用跨语言特征训练的维语声学模型的鲁棒性。以测试对包含信道、噪音、口音、情绪等多种因素的语音声学特征在哪种声学模型上更具区分性。

2. 声学模型训练技术

语音识别就是将观测到的语音样本O通过函数$\mathcal{F}(\cdot)$转换成文本$\mathcal{T}=\mathcal{F}(O)$的过程。其核心问题是寻找适合的转换函数$\mathcal{F}$，使得转换得到的文本与人工标注尽可能接近。描述自动语音识别系统输出与人工

标注之间相似度的指标通常是字错误率(Word Error Rate, WER)。系统的开发目标就是要得到一个在识别时能够产生最小字错误率的转换函数。语音识别系统结构示意图如图 1 所示。

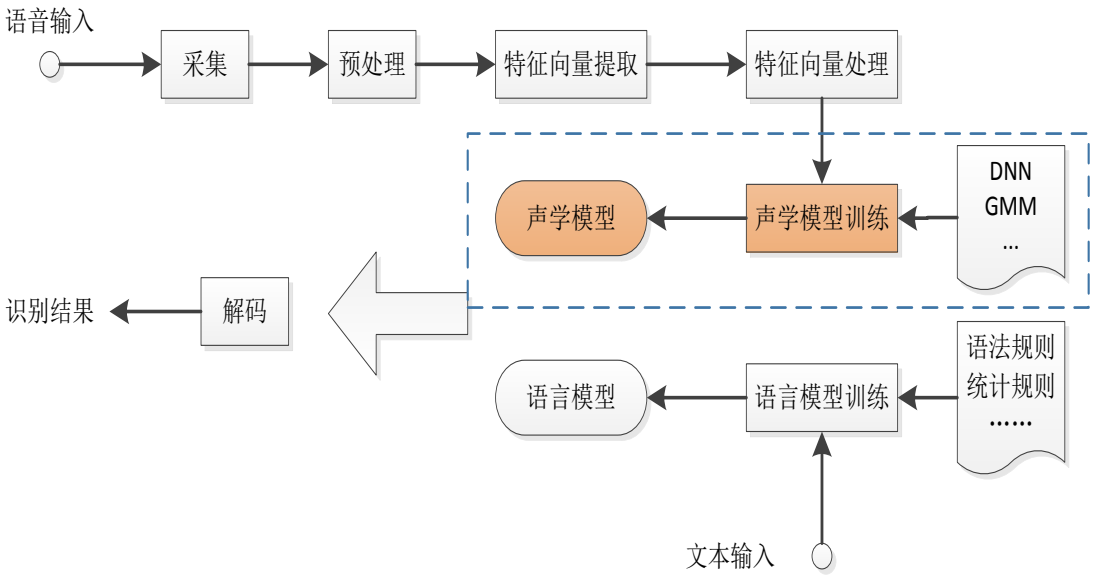

图 1　语音识别系统架构

语音识别问题是非常复杂的，我们不可能利用手工指定规则或使用专家知识轻易的得到合适的转换函数。通常的做法是利用统计模型在训练数据中寻找规律，进行统计语音识别。设某一个词序列表示为$W = \{w_1, w_2, \cdots, w_N\}$，在统计语音识别框架下评价其与某一段观测$O$匹配程度的度量通常使用如下的贝叶斯公式：

$$p(W|O) = \frac{p(O|W)p(W)}{p(O)} \tag{1}$$

其中，$p(O|W)$被称为声学模型概率，表示语音本身的声学特征与词串W的匹配程度。通常，我们使用隐马尔科夫模型(Hidden Markov Model, HMM)来对声学模型进行建模。$p(W)$则是语言模型概率，表示在自然语言中词串W本身可能出现的概率。贝叶斯决策理论就是选择

使得上式最大的词串W^*作为自动语音识别的输出。

2.1 通用声学模型训练技术

声学模型的目标是提供一种有效的方法，计算语音的特征矢量序列和每个发音模板之间的距离。声学模型的设计和语言发音特点密切相关。模型识别单元大小(词发音模型、字发音模型、半音节模型或音素模型)对语音训练数据量大小、语音识别率，以及灵活性有较大的影响。对中等词汇量以上的语音识别系统来说，识别单元小，则计算量也小，所需的模型存储量也小，要求的训练数据量相对也小，所需的模型存储量也小，要求的训练数据量相对也少，但带来的问题是对应语音段的定位和分割困难，以及更复杂的识别模型规则。通常大的识别单元易于包括协同发音在模型中，这有利于提高系统的识别率，但要求的训练数据相对增加。因此，需要根据语音的特点选择建模单元，并为不同的建模单元建立不同的 HMM。

语音识别建模单元的选取应该考虑到一致性、可训练性和可共享性。所谓一致性，是指不同语音实例中相同的语音单元需要具备声学上一致的特征；可训练性是指对于一个建模单元需要得到足够的训练数据来对其参数进行估计；而可共享性则是指是否可以在不同的建模单元之间共享某些具有共性的训练数据。

正是基于上述 3 点，通常的语音识别建模单元一般采用音素(phone)、音节(syllable)以及词(word)等。而在大词汇量连续语音识别中，常用音素作为基本的建模单元。此外，为了考虑连续语音中的协同发音(Coarticulation)现象，一般还采用上下文相关

(Context-dependent)的音素建模，如三元音素(Tri-phone)建模等。

在建模单元确定以后，我们就可以根据其特点为各单元分配适当的 HMM 拓扑结构。一般来说，对普通的音素单元常采用自左向右的无跨越 HMM，而对静音模型 sil 及短停顿模型 sp 等，则可采用可跨越的 3 状态及单状态 HMM，如图 2 所示。

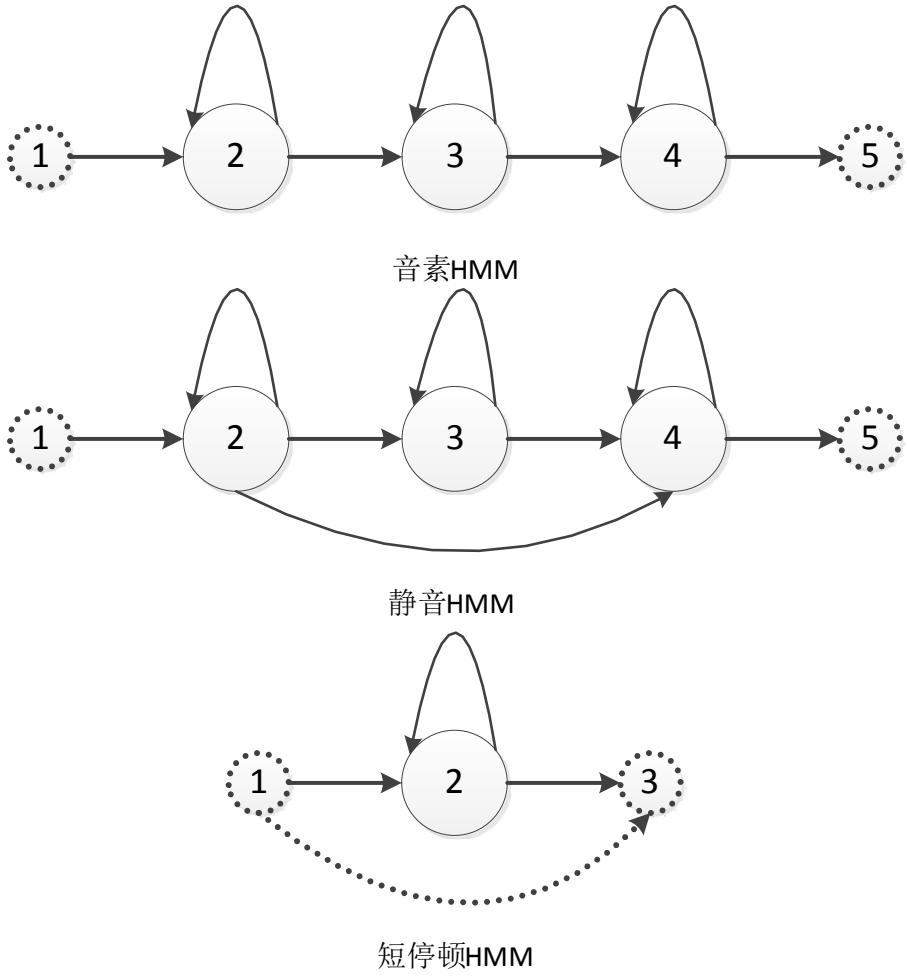

图 2 不同建模单元对应的 HMM 拓扑结构

对于每个 HMM 状态的输出概率，在大词汇量连续语音识别任务中通常用高斯混合模型(Gaussian Mixture Model，GMM)或深度神经网络(Deep Neural Networks, DNNs)来表达，其中，GMM 表达如下：

$$b_i(x) = \sum_{m=1}^{M} \frac{c_{im}}{\sqrt{(2\pi)^D |\sum_{im}|}} \exp\left[-\frac{1}{2}(\mathrm{x} - \mu_{im})^{\mathrm{T}} \sum_{im}^{-1} (\mathrm{x} - \mu_{im})\right] \quad (2)$$

其中，D是观测向量x的维度，μ_{im}和$\sum_{im}$则分别为状态i中第m个高斯混合的均值向量及协方差矩阵，c_{im}是该混合高斯成分的权重。对于一个典型的语音识别系统，每个状态的混合高斯成份数至少需要 8 到 12 个，因此也就对训练数据量的规模提出了更高的要求。同时应当看到，声学模型建模精度与训练数据量的矛盾在目前仍是极为突出的：为了更精细的建模，我们必须采用上下文相关模型，并使用更多的混合高斯成份数。这样一来，就使得模型参数的数目急剧增加。而与之相对，训练数据相比语音中的复杂现象来说，本来就非常稀疏，要在如此稀疏的数据之上可靠地估计如此之多的模型参数又会成为另一个困难的任务。有鉴于此，声学模型参数绑定技术就显得尤为重要。目前，在大词汇量连续语音识别中最常用的模型参数绑定方法是最大似然准则下基于决策树(Decision Tree)的状态绑定。除此之外，我们还可以在模型、状态、混合高斯成份乃至特征等多个层面对模型参数进行绑定与压缩。这些技术确保了模型参数在较高的建模精度下仍能得到可靠的估计。

长期以来，语音识别系统在描述每个建模单元的统计概率模型时，大多采用的是 GMM。这种模型由于估计简单，适合海量数据训练，同时有成熟的区分性训练技术支持，长期以来，一直在语音识别应用中占有垄断性地位。但这种混合高斯模型本质上是一种浅层网络建模，不能充分描述特征的状态空间分布。另外，GMM 建模的特征维数一

般是几十维，不能充分描述特征之间的相关性。最后，GMM 建模本质上是一种似然概率建模，虽然区分度训练能够模拟一些模式类之间的区分性，但区分能力有限。随着深度学习技术在机器学习领域的发展，基于深度学习的深度神经网络(DNNs)技术也彻底改变了语音识别原有的技术框架。采用深度神经网络后，可以充分描述特征之间的相关性，可以把连续多帧的语音特征并在一起，构成一个高维特征。最终的深度神经网络可以采用高维特征训练来模拟。由于深度神经网络采用模拟人脑的多层结果，可以逐级地进行信息特征抽取，最终形成适合模式分类的较理想特征。这种多层结构和人脑处理语音信息时，是有很大的相似性的。深度神经网络的建模技术，在实际线上服务时，能够无缝地和传统的语音识别技术相结合，在不引起任何系统额外耗费情况下，大幅度提升了语音识别系统的识别率。其在线的使用方法具体如下：在实际解码过程中，声学模型仍然是采用传统的 HMM 模型，语言模型仍然是采用传统的统计语言模型，解码器仍然是采用传统的动态 WFST 解码器。但在声学模型的输出概率计算时，用神经网络的输出后验概率乘以一个先验概率来代替传统 HMM 模型中的 GMM 的输出似然概率。

由于深度神经网络的多隐藏层结构，使得 DNN 具有从原始数据中学习层次特征的能力，因此 DNN 在表征语音信号复杂模式上具有高度灵活性，这使得 DNN-HMM 模型在语音识别系统声学模型建模中取得了巨大的成功，全面代替传统的 GMM-HMM 模型。图 3 为 GMM 模型与 DNN 模型，图 4 为 DNN-HMM 模型框架图。

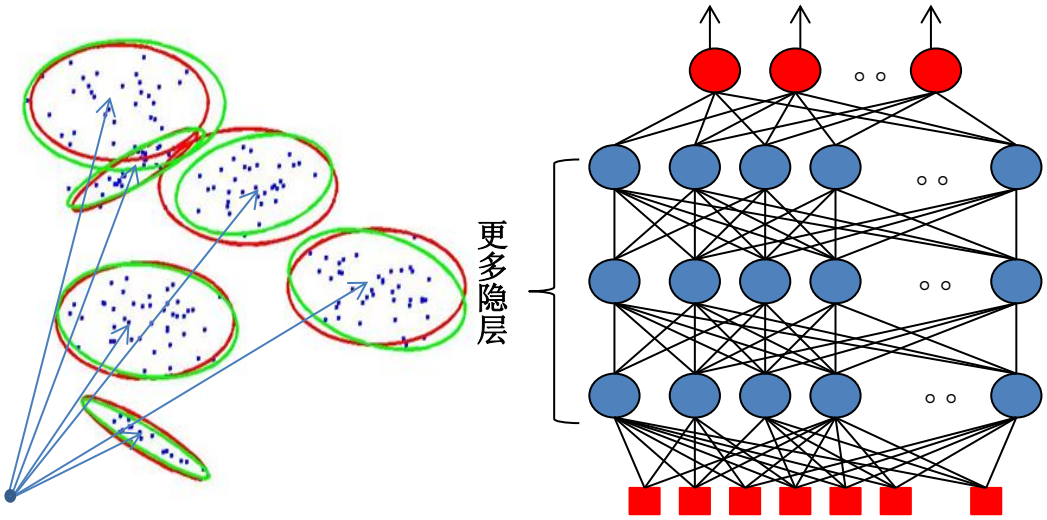

图 3　GMM 模型和 DNN 模型

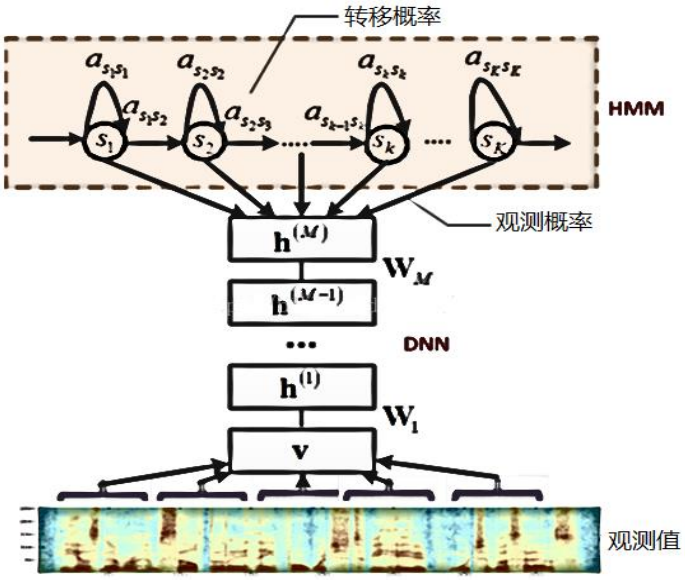

图 4　DNN-HMM 模型框架图

由图 3 可以看出，相比较于传统的 GMM 模型只有均值、方差两个参数，DNN 模型具有多隐藏层（通常大于 3 层）的网络结构，并且每一层都有许多节点（通常上千个），相邻层次间是通过权值 W 全连接的，并且每个隐藏层的输出都经过非线性映射后作为下一层的输入。

结合图4可发现HMM模型中的发射概率是通过DNN模型来计算的，而 HMM 模型的转移概率为 DNN 模型提供原型，即音素与共享状态的对应。

从上述分析中可以看出， DNN 模型的本质思想是堆叠多个神经元层，每个层都提取一定的特征和信息，以此来模仿人脑机制来解释数据，并且，训练过程中，只需指定网络的层数，而不需要给定具体的参数，网络通过计算数据来自动学习最终的参数，不一样的网络参数能够识别不同的物体，训练好的网络就能自动识别物体。因此当数据量过于稀疏时，网络无法学习出语音信号中各种复杂的特征，而导致训练所得声学模型失配。

2.2 基于跨语言特征的低资源维语声学模型训练技术

根据上文所述可知，传统语音声学模型训练方法中，用于计算声学模型的输出分布时，GMM 受限于本身似然估计的性质，不能表现出强劲的区分性；而具有高维特征表征能力，能充分描述特征之间的相关性，可以模拟人脑来区分复杂语音信号的 DNN 网络结构，却又因为其网络结构的复杂、庞大，不适合在数据量稀疏的情况下训练语音声学模型。

本文提出基于跨语言特征的低资源维语声学模型训练技术，即是充分利用 DNN 网络的特性，并且借鉴语言在最基本的音素单元上有共性，能实现共享的特点，以原始大规模的汉语语料为基础，训练以

维语为目标的语音声学模型，解决由于维语语料稀疏所带来的声学模型失配的问题。

一个 DNN 模型可以简单表示为一个三元组(X,W,K,Y)，其中 $X=(x_1,x_2,x_3\cdots x_n)$为输入向量集合，$W=(w_{j1},w_{j2},w_{j3}\cdots w_{jn})$为连接权值集合，$K=(1,2,3\cdots k)$为隐藏层数，$Y=(y_1,y_2,y_3\cdots y_m)$为输出向量集合。因此一个稳定的声学 DNN 模型，即通过大量 X 训练 W。图 5 所示为训练充分 DNN 抽象识别过程。

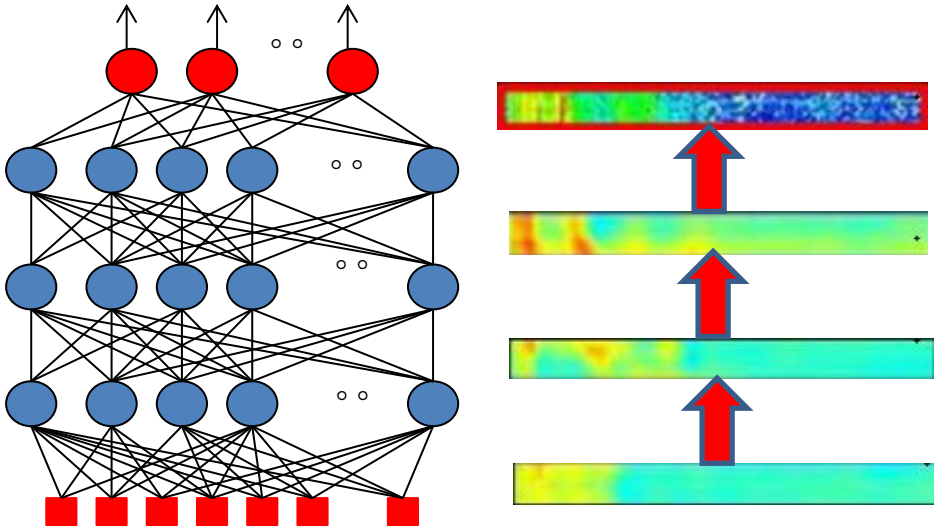

图 5　训练充分 DNN 模型抽象识别过程

通过 DNN，我们可以把多语言的 DNN 模型结合起来，形成一个底部训练充分，抽象特征可供顶部共享的自适应模型，从而解决训练数据过于稀疏导致训练不充分，模型失配的问题。简单而又鲁棒。如图 6，即是通过汉语声学模型构建的自适应维语声学模型。

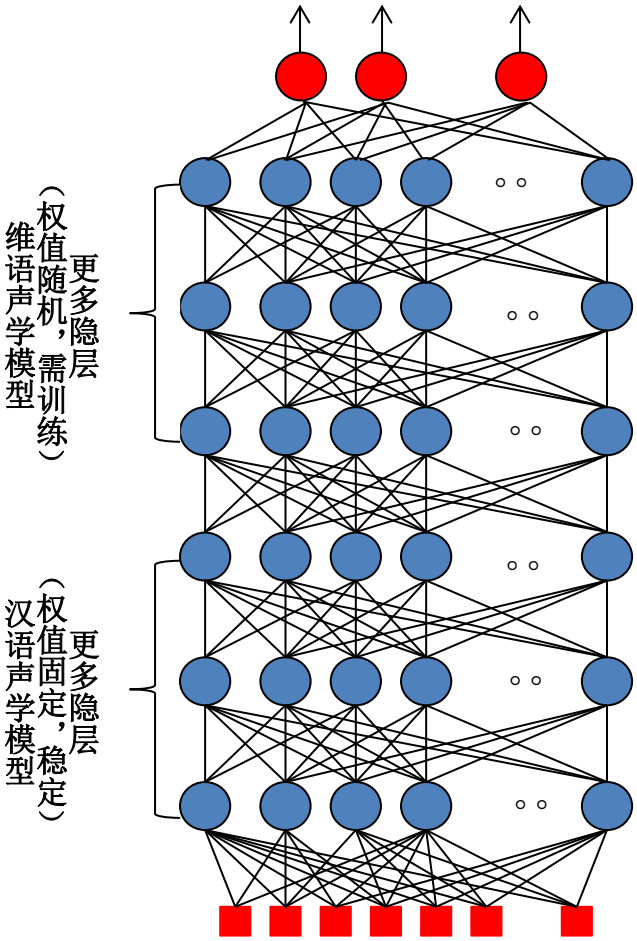

图 6　<汉语>+<维语>自适应声学模型

结合 DNN 模型的学习过程——正向传播、反向传播，本文将进一步介绍基于跨语言特征的低资源维语声学模型训练技术的优越性。

2.2.1 正向传播

令o_i=前层神经元的输出，设第K层（有H_k个神经元）输入整合

$net_j = \sum_{i=0}^{H_k} w_{ij} o_i$，神经元$j$的输出（激活）$o_j = f(net_j) = \frac{1}{1+e^{-net_j}}$，即有$L$个隐层的 DNN 网络满足：

$$y_k = \sum_i w_{k,i}^{(L+1,L)} f_i^{(L)}(\sum_j w_{j,m}^{(L,L-1)} f_m^{(L-1)}(\cdots \sum_q w_{p,q}^{(2,1)} f_q^{(1)}(\sum_r w_{q,r}^{(1,0)} x_r))) \quad (3)$$

其中，y_k为输出，$f(\cdot)$为 sigmoid 函数，$\left\{w_{i,j}^{(l,l-1)}\right\}_{i,j}$表示连接第$l-1$层$i$个神经元与第$l$层$j$个神经元之间的权值；$\{x_r\}_r$表示输入的特征向量。

2.2.1 反向传播

采用δ学习算法，调整各层间的权值，使误差函数E最小。

1.误差函数的可调参数个数n_w等于各层权值数加上阈值数，即：

$$n_w = \sum_{i=0}^{L} H_{i+1}(H_i + 1)$$

2.误差E是$n_w + 1$维空间曲面，曲面上每个点的“高度”对应于一个误差值，每个点坐标向量对应着n_w个权值，因此称为误差的曲面空间。见图 7。

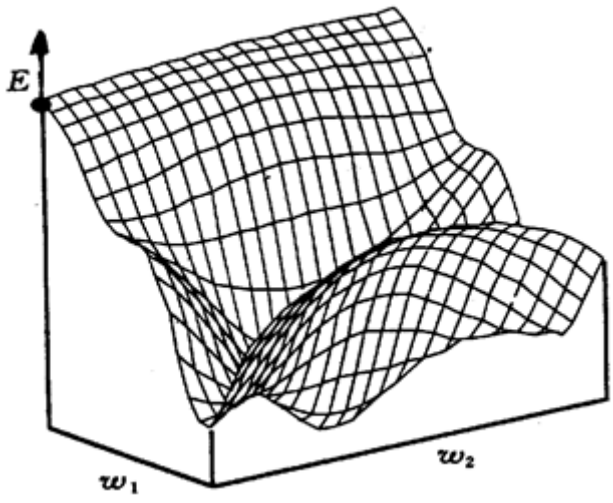

图 7 误差曲面

3.训练过程中采用随机梯度下降(stochastic gradient descendent, SGD)算法。

一个典型的机器学习的过程，首先给出一组输入数据X，算法会通过一系列的过程得到一个估计的函数，这个函数有能力对没有见过的新数据给出一个新的估计Y，也被称为构建一个模型。例如用x_1、$x_2 \ldots x_n$去描述特征分量，用$h(x)$来描述我们的估计，得到模型如下：$h(x) = h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$，因此，可以引入误差函数：$\mathrm{E}(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$，目标函数为：$\min_\theta \mathrm{E}_\theta$，即调整$\theta$使误差最小化。

假设随机站在E函数构成的曲面上的一点，要以最快的速度到达最低点处，即应该沿着坡度最大的方向向下走（梯度的反方向）。用数学公式描述如下：

$$\frac{\partial}{\partial\theta}\mathrm{E}(\theta) = \frac{\partial}{\partial\theta}\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x) - y)^2 = (h_\theta(x) - y)x^{(i)} \qquad (4)$$

根据公式(4)可以推出θ的更新过程如下（α表示学习速率）：

$$\theta_i = \theta_i - \alpha\frac{\partial}{\partial\theta}\mathrm{E}(\theta) = \theta_i - \alpha(h_\theta(x) - y)x^{(i)} \qquad (5)$$

通过图 6 的分析，结合式(3)(4)(5)，不难发现，基于跨语言特征的低资源维语声学模型训练技术即是在前向过程中，由训练充分的声学模型（汉语、英语等），提供低层的共享信息，即(3)式中部分$w_{i,j}$值已经基本固定，加入少量维语数据后，在反向传播过程中，根据目标值与实际值的误差，按照(5)式的过程微调网络即可得到一个稳定的声学模型。由此可见，基于跨语言特征的低资源维语声学模型训练技术相对于传统的声学模型训练技术，不会出现由于训练数据过于稀疏而

导致的模型失配问题，并且在训练速度上也将得到大幅的提高。

3. 基于跨语言特征的低资源维语声学模型训练技术实现

本文所述基于跨语言特征的低资源维语声学模型训练技术是在 kaldi 平台上实现的。以训练充分的汉语 DNN 声学模型为源模型，然后辅以少量维语数据构建的 DNN 声学模型为目的模型，来构建稳定的维语声学模型。具体实现如下：

(1)安装 kaldi；详见：

/nfs/disk/work/users/zhangzy/work/kaldi-201311

(2)训练充分的汉语 DNN 声学模型；详见：

/work0/yinshi/dnn_model_train/uyghur_fbank_train/exp/tri4b_dnn_uyghur_fbank_based-6000h-mpe-DNN_only-classify_merge/nnet.init

(3)以少量维语数据训练 DNN 模型；详见：

/work0/yinshi/dnn_model_train/uyghur_fbank_train/exp/tri4b_dnn_uyghur_fbank_mpe/4.nnet

(4)通过 nnet-copy 裁剪汉语 DNN 模型，保留低部层次，裁剪掉输出及后面一些层次；通过 nnet-copy 裁剪维语 DNN 模型，保留输出层次，裁剪掉前面的层次;通过 nnet-concat 连接两个经过裁剪之后的模型；详见：

/work0/yinshi/dnn_model_train/uyghur_fbank_train/exp/tri4b_dnn_uyghur_fbank_based-6000h-mpe-DNN_only-classify_merge/nnet.init

(5)将连接后的模型作为新声学模型的 MLP 初始模型，替代原本的随

机初始化模型。即设置：

--mlp_init=/work0/yinshi/dnn_model_train/uyghur_fbank_train/exp/tri4b_dnn_uyghur_fbank_based-6000h-mpe-DNN_only-classify_merge/nnet.init --hid-layers

(6)将上述配置用于训练维语声学模型。

4. 总结

本文从理论及实现两方面详述了基于跨语言特征的低资源维语声学模型训练技术的特点，并且对比分析了这种声学模型训练技术与传统声学模型训练技术的异同，更为鲜明的突出了该技术的优势。实验表明该种技术不仅可以解决训练数据稀疏导致的模型失配问题，还能减少训练时间，提高效率。

(实验结果详见：

http://cslt.riit.tsinghua.edu.cn/cgi-bin/cvss/cvss_request.pl?account=yinshi&step=view_request&cvssid=274)