ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 2,Mar. 2017,pp. 205－220
DOI:10.16337/j.1004-9037.2017.02.001

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

低资源语音识别若干关键技术研究进展

刘　加　张卫强

(清华大学电子工程系,北京,100084)

摘　要:低资源语音识别是当今语音界研究的热点问题之一,也是多语言小语种语音识别技术在实际应用中所面临的重要挑战之一。本文回顾并总结了低资源语音识别的发展历史和研究现状,重点介绍了低资源语音识别在声学特征、声学模型和语言模型方面的若干关键技术研究进展。具体内容包括发音特征、多语言瓶颈特征、子空间高斯混合模型、卷积神经网络声学模型和递归神经网络语言模型,然后介绍了针对低资源语音识别的公开关键词搜索(Open keyword search,OpenKWS)评测,最后对低资源语音识别进行了总结和展望。

关键词:语音识别;低资源;声学模型;语言模型

中图分类号:TP319　　**文献标志码**:A

Research Progress on Key Technologies of Low Resource Speech Recognition

Liu Jia, Zhang Weiqiang

(Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China)

Abstract: Low resource speech recognition is one of currently researching hotspots in speech recognition community, and is also one of the important challenges for the application of multilingual and minority language speech recognition technologies. This paper summarizes and reviews the current states and history of low resource speech recognition, and introduces several key technologies, including articulatory feature, multilingual bottleneck feature, subspace Gaussian mixture model, convolutional neural network based acoustic model and recurrent neural network based language model. After that the open keyword search (OpenKWS) evaluation is introduced. Finally, the prospective of low resource speech recognition is presented.

Key words: speech recognition; low resource; acoustic model; language model

引　　言

语音是人类最主要通信手段之一,语音信号一直以来是信息获取的主要来源之一[1,2]。无论是民用还是军事,一直以来都受到世界各国的高度重视。近年来,随着互联网技术迅猛发展,各种音频信息量爆炸性地迅速增长,基于网络的多语言语音识别技术应用需求越来越迫切。近几年来主要大语种(汉语和英语)的语音处理技术(如语音识别技术、语种识别技术、说话人识别技术、关键词检测技术和语音合成技术等)已经在互联网、手机、呼叫中心及信息安全领域中开始得到应用。一方面,现有语音识别系统

基金项目:国家自然科学基金(61370034,61403224)资助项目。
收稿日期:2015-06-05;**修订日期**:2015-06-30

的准确率和稳健性还不够好，有待于进一步提高，尤其是针对信道质量较差的电话语音和噪声语音，识别性能会急剧降低，不能满足应用需求，这需要从建模方法和理论上进一步深化研究；另一方面，随着网络中多语言应用越来越普及，中国现有的以汉语与英语为主的语音识别系统已经无法应对众多方言、少数民族语音及其他小语种语音。同时现有的识别系统对处理多语言语音混用情况下的识别和口语化语音的识别也存在技术上的困难，未得到很好的解决。众所周知，目前的语音识别系统是建立在大量训练数据基础之上的，一旦训练数据缺失，将出现“巧妇难为无米之炊”的情形，原本成熟的方法很难发挥出应有的水平，甚至根本无法工作。低资源语音识别正是针对这种情况进行研究突破的，国际上一般称为“Low resource speech recognition”。低资源语音识别是用于训练的数据资源匮乏甚至缺失，这些数据资源包括语音、标注、发音字典和文本，某一方面或几方面的资源匮乏都属于低资源语音识别。

目前，国际上低资源语音识别研究非常活跃[3-35]。作者在 Ei 数据库[36]对低资源语音识别进行了文献检索，按年份统计的结果如图 1 所示。从论文发表数量来看，在总体上呈逐年增加的趋势，这在一定程度上反映了低资源语音识别的研究逐年升温。

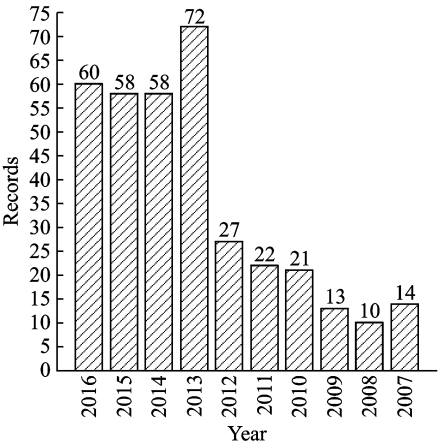

图 1 Ei 数据库搜索关于 Low resource 和 Speech recognition 论文按年份统计结果（截至日期：2016 年 12 月 31 日）

Fig. 1 Distribution of papers on low resource and speech recognition according to year in Ei database (Statistics by Dec. 31, 2016)

2009 年，由来自美国、德国和捷克等国家和地区的语音语言学专家所组成的研究小组在美国约翰霍普金斯大学举行了主题为“低资源数据条件下新语种的低成本、高质量语音识别”的专项研讨会[37]，专门针对那些使用人口少、训练数据获取困难及语言、语音专家稀缺的语言种类的语音识别技术展开研究，这为低资源语音识别奠定了基础。2011 年初，美国情报高等计划研究署（IARPA）开始实行 Bable 计划[38,39]，该计划最终由美国和欧洲多个研究机构承担，主要研究语音识别中低数据资源的语音识别系统快速构建。Bable 计划的目标是：在当前常用语言识别系统的基础之上，仅需要一周的时间，即可以完成对任何其他语言的语音识别系统的构建，大量减少系统构建的资源开销，极大加快系统构建周期。参加该计划的团队来自世界上语音方面最领先的单位，如：MIT，Cambridge，CMU，JHU，IBM，SRI 和 BBN 等。2015 年本团队参加了 Babel 计划中的公开关键词搜索（Open keyword search，OpenKWS）评测，并取得公开测试条件第一名的好成绩。

本文介绍了低资源语音识别方面的若干典型工作和关键技术研究进展，包括发音特征、多语言瓶颈特征、子空间高斯混合模型、卷积神经网络声学模型和神经网络语言模型，这些工作是作者近年来在一线从事科研工作的总结；之后介绍了 OpenKWS 评测，这是针对低资源语音识别的目前最具影响力的一个国际学术评测。

1 典型语音识别系统架构[40-43]

在现代语音识别框架下，语音识别问题可以抽象地表示为一个最优化问题[44]。其数学原理是在给定观测特征矢量序列 $\boldsymbol{O}$ 的情况下，求解最优词序列 $\hat{\boldsymbol{W}}$，使得条件概率 $p(\boldsymbol{W}|\boldsymbol{O})$ 最大化。根据贝叶斯公式，问题可以进一步表示为

$$\hat{\boldsymbol{W}}=\arg\max_{\boldsymbol{W}} p(\boldsymbol{W}\mid\boldsymbol{O})=\arg\max_{\boldsymbol{W}}\frac{p(\boldsymbol{O}\mid\boldsymbol{W})\cdot p(\boldsymbol{W})}{p(\boldsymbol{O})}=\arg\max_{\boldsymbol{W}} p(\boldsymbol{O}\mid\boldsymbol{W})\cdot p(\boldsymbol{W}) \tag{1}$$

式中：概率 $p(\boldsymbol{O})$ 对于给定的词序列 $\boldsymbol{O}$ 固定，它不影响识别结果 $\hat{\boldsymbol{W}}$，因而可以忽略。概率 $p(\boldsymbol{O}|\boldsymbol{W})$ 代表观

测矢量 $\boldsymbol{O}$ 对给定词序列 $\boldsymbol{W}$ 的似然度,这就是声学模型的建模任务。而概率 $p(\boldsymbol{W})$ 建模了词序列出现的先验概率,它依靠语言模型来进行估计。在大词汇量连续语音识别系统中,词序列 $\boldsymbol{W}$ 会依靠发音字典进一步展开成上下文无关的音素序列 $\boldsymbol{L}$。引入上下文关系后,上下文无关的音素序列 $\boldsymbol{L}$ 还可以展开成上下文相关的音素序列 $\boldsymbol{C}$,而序列 $\boldsymbol{C}$ 最终展开成隐马尔科夫模型(Hidden Markov model,HMM)的状态序列 $\boldsymbol{S}$。将语音识别的过程概括成统一的公式,即有

$$\hat{\boldsymbol{W}}=\arg\max_{\boldsymbol{W}}\sum_{\boldsymbol{L}}\sum_{\boldsymbol{C}}\sum_{\boldsymbol{H}}p(\boldsymbol{O}\mid\boldsymbol{S})p(\boldsymbol{S}\mid\boldsymbol{C})p(\boldsymbol{C}\mid\boldsymbol{L})p(\boldsymbol{L}\mid\boldsymbol{W})p(\boldsymbol{W}) \tag{2}$$

从式(2)可以看出,大词汇量连续语音识别包含了 5 层映射关系,其中 $p(\boldsymbol{L}|\boldsymbol{W})$ 由发音字典决定,$p(\boldsymbol{C}|\boldsymbol{L})$ 通常将单音素映射成上下文相关的三音素,$p(\boldsymbol{S}|\boldsymbol{C})$ 由声学模型的状态聚类来决定,这 3 个概率都并不与数据直接相关。因此声学模型归根到底建模的是观测矢量 $\boldsymbol{O}$ 到 HMM 的状态 $\boldsymbol{S}$ 的似然度 $p(\boldsymbol{O}|\boldsymbol{S})$。

根据语音识别的基本原理,可以进一步给出语音识别系统的基本组成框图,如图 2 所示。在语音识别系统中,声学模型由大量带标注的语音训练得到,语言模型由大量文本训练得到。解码器综合了发音字典、声学模型和语言模型等知识源,对测试语音的特征进行识别,得到最终的文本识别结果。

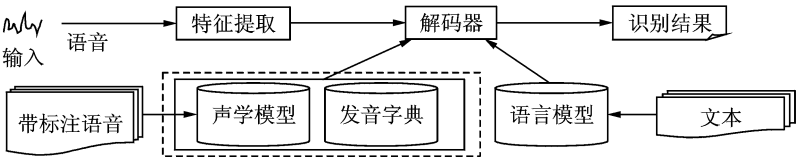

图 2　语音识别系统基本组成

Fig. 2　Basic flowchart of speech recognition system

2　低资源语音识别中的声学特征

2.1　语言无关的发音特征[41]

发音特征,在一般文献报道中翻译为“Articulatory feature”,指人类在产生声音时唇和舌的位置、形态等信息,比如平常说的摩擦音、鼻音等。国际语音协会将人类语言按照发音特征进行分类,每一类发音均可以表示成一系列发音特征的组合,这每一类发音就称为音素。每一个音素都是一组特定发音特征的组合,同样每一种发音特征又同时存在于很多音素的发音过程中。以表 1 的英语为例,左边两列是经常可以看到的根据发音特征的音素分类;反之,也可以得到像右边两列所示的音素和发音特征相对应的表。从表 1 中可见,发音特征与音素有着千丝万缕的联系,可以说是“你中有我,我中有你”。

表 1　音素与发音特征对应关系示例

Tab. 1　Relation between phonemes and articulatory features

发音特征	音素	音素	发音特征
Fricative	jh ch s sh z f th v dh	f	Fricative labial
Nasal	m n ng	th	Fricative dental
Labial	b f m p v w	aa	Vowel low back
⋮	⋮	⋮	⋮

考虑到音素建模在语音识别中的重要地位以及音素与发音特征之间的联系,很多学者尝试从发音特征层面寻找各种建模的方式,希望可以找到另一条通用语音识别声学建模大门。许多研究者都对发音特征进行了详细深入的研究[45,46],同时一些基于发音特征的语音识别系统也取得了不错的效果[47,48];但由于发音特征相比音素,分类标准不确定度大,分类边界不够清晰以及离散化的表示不利于认知等缺

点，基于它的研究还只处于初期阶段，基于发音特征建立起来的语音识别系统相比传统系统来说性能稍差。发音特征相比音素是更加具有普遍意义的分类集合。不仅同一语言中的不同音素通过一系列发音特征进行划分；不同语言的各种音素也是通过这些发音特征进行划分，因为无论是说哪一种语言，人类在产生声音时用于描述唇、舌的位置与形态的状态量就只是固定的若干种而已。发音特征相比与音素，是在各语言中差异性更加小的声学单元。可以从平均共享因子的分布特性来进一步说明发音特征的语言通用性[49]。平均共享因子是指多种语言混合时共享声学单元的平均语言数量，比如 b 在汉、英和日中都出现，那么 b 的共享因子就是 3。根据共享单元的区别，可以分为音素平均共享因子和发音特征平均共享因子两类。图 3 所示为一个平均共享因子随混合语言数变化曲线的示意图，利用基于 IPA 国际音标体系表示下的汉、英、日、西和德 5 种语言来进行两类平均共享因子的统计。从图 3 可以看到，发音特征平均共享因子随着混合语言数量呈线性增长，而音素平均共享因子很快趋于饱和，这说明发音特征在不同语言中的表现都是一致的，各种不同语言的发音特征基本一样；但是不同语言的音素单元表示千差万别，不同语言所用音素集合有很大的不同，所以发音特征相比音素，是在各语言间差异性更小的共性单元。

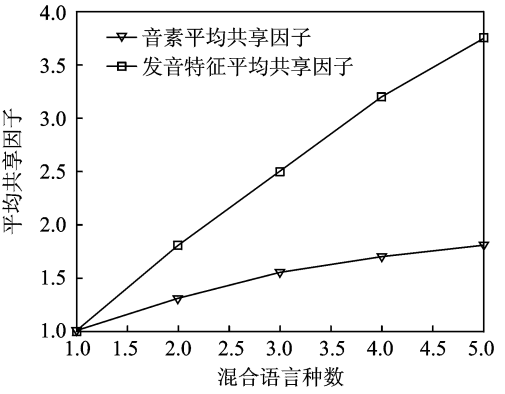

图 3 音素与发音特征平均共享因子对比图

Fig. 3 Comparison between average sharing factor of phonemes and articulatory features

基于以上对发音特征单元的分析，从发音特征层面入手来发挥它在多语言通用性上的优势，从而解决低数据资源条件下的语音识别问题[50-52]：(1) 发音特征的语言无关性使得它在不同语言间的差异很小，和新加入语言的差异很小；(2) 发音特征的高共享因子说明它可以充分利用不同语言训练数据来进行共享训练，不会出现模型训练不充分、数据稀疏等问题；(3) 基于发音特征的建模方法相比传统方法是一个全新的尝试，从一个新层次来进行语音建模，与传统基于音素建模的方法有一定的互补优势。

2.2 多语言数据共享的瓶颈特征[43]

当前深度神经网络(Deep neural network，DNN)在语音识别中的应用主要有两种：第一种是混合(hybrid)方法，即利用 DNN 直接估计 HMM 模型的物理状态后验概率，本文之前的研究都属于该方法的范畴；而第二种是级联(tandem)方法，即利用 DNN 进行特征提取，再利用高斯混合模型-隐马尔科夫模型(Gaussian mixture model-hidden Markov model，GMM-HMM)声学模型或深度神经网络-隐马尔科夫模型(Deep neural network-hidden Markov model，DNN-HMM)声学模型对提取出的特征进行建模。由于 hybrid 方法结构简洁、性能可靠，故该方法是目前基于 DNN 的声学建模中较为主流的方法。而 tandem 方法在某些特定的应用场合却有 hybrid 方法无法比拟的优点，例如：(1)该方法能够利用 GMM-HMM 模型的一些优势，如 GMM-HMM 模型能够很方便地使用计算机集群进行并行训练，当使用相对少量数据训练 DNN 来进行特征提取，而使用大量数据训练 GMM-HMM 模型时，可以加速模型训练过程，增强模型在大数据条件下的性能[53]。(2)在 hybrid 方法中，由于 DNN 直接起到声学建模的作用，因此方法要求 DNN 的训练数据和测试数据相匹配。在 tandem 方法中，用来训练后端声学模型的训练数据仍然需要和测试数据匹配，但是由于 DNN 仅作为前端特征提取器，此时 DNN 的训练数据和测试数据的匹配性要求并不很高，且 DNN 特征提取器的训练数据未必和后端声学模型的训练数据相同。这一性质为跨语言声学建模任务提供了一条可行的思路。

在 tandem 建模方法中，利用神经网络进行特征提取的方法又可进一步分为两种：第一种是利用神经网络的输出层提取特征。由于神经网络的输出具有概率意义，该特征也被称为“概率特征”(probabilistic feature)[54]。第二种是利用神经网络的隐含层提取特征。由于用来进行特征输出的隐含层节点数

一般较少,构成了一个神经网络的瓶颈结构,因此该特征被称为"瓶颈特征"(bottleneck feature)[55]。在DNN出现之前,概率特征较为常见。而DNN出现之后,大多数tandem方法使用bottleneck特征,其原因是bottleneck特征的维数可以灵活配置,且此时DNN的输出层可以使用大量的节点以提高性能。而概率特征输出维数固定,一般还需要经过特征变换(如主成分分析)降维后才可进行声学建模。

Bottleneck特征提取器示意图如图4所示。用于提取bottleneck特征的DNN同hybrid方法中的DNN具有非常类似的结构,其输出层同样对应HMM的物理状态。因此,bottleneck特征中包含了丰富的语音音素信息,正是该信息能够带来后续声学模型性能的提升。在提取bottleneck特征所用的DNN中,某个隐含层的节点数较少,该节点的神经元激活函数输出即为bottleneck特征。当提取bottleneck所用的DNN模型训练完成后,其bottleneck层之后的隐含层和输出层参数并不再被使用,因此最终的bottleneck特征提取器仅包括DNN从输入层到bottleneck层之间的网络参数,如图4中的虚线部分所示。

Bottleneck特征既可用于训练的GMM-HMM模型,也可用于训练DNN-HMM模型。当利用bottleneck特征训练DNN-HMM模型时,由于DNN-HMM的输入为若干帧相邻的特征,因此从基本特征输入到声学模型输出的整个网络结构等价于一个时域上的卷积神经网络,该结构如图4所示。与时域上的卷积神经网络不同的是,此网络在训练过程中bottleneck特征提取部分的参数并不再变化,仅bottleneck特征之上的部分利用有标注数据进行训练。

在bottleneck层神经元类型的选择上,通常并不使用sigmoid函数,而是使用线性函数。将bottleneck层记为第l层,当该层使用线性激活函数,而其他层使用sigmoid非线性函数时,有

$$\boldsymbol{y}^{l+1}=\theta(\boldsymbol{W}_{l+1}^{\mathrm{T}}\cdot(\boldsymbol{W}_{l}^{\mathrm{T}}\cdot\boldsymbol{y}^{l-1}+\boldsymbol{b}_{l})+\boldsymbol{b}_{l+1})=\theta((\boldsymbol{W}_{l+1}^{\mathrm{T}}\boldsymbol{W}_{l}^{\mathrm{T}})\cdot\boldsymbol{y}^{l-1}+(\boldsymbol{W}_{l+1}^{\mathrm{T}}\boldsymbol{b}_{l}+\boldsymbol{b}_{l+1})) \tag{3}$$

式中:θ为非线性函数,$\boldsymbol{y}^{l+1}$为第$l+1$层的输出,$\boldsymbol{W}_{l+1}$和$\boldsymbol{b}_{l+1}$分别为第$l+1$层的权重矩阵和偏置矢量。从式(3)可以看出,当bottleneck层使用线性神经元函数时,第$l+1$层的输出$\boldsymbol{y}^{l+1}$和第$l-1$层的输出$\boldsymbol{y}^{l-1}$之间的关系类似于普通DNN相邻隐含层之间的关系,其等价的权重矩阵为$\boldsymbol{W}_{l+1}^{\mathrm{T}}\boldsymbol{W}_{l}^{\mathrm{T}}$,等价的偏置矢量为$\boldsymbol{W}_{l+1}^{\mathrm{T}}\boldsymbol{b}_{l}+\boldsymbol{b}_{l+1}$。此时,bottleneck层较少的神经元节点仅影响等价权重矩阵$\boldsymbol{W}_{l+1}^{\mathrm{T}}\boldsymbol{W}_{l}^{\mathrm{T}}$的秩,DNN的权重矩阵往往为低秩矩阵[56,57]。因此线性bottleneck层对DNN模型的精度影响最小,换言之,bottleneck层能够保留的信息也最多。

为了能够在低资源、多语言的现实环境中更好地利用多个非目标语言的数据,以辅助目标语言的声学建模,采用基于多语言bottleneck特征的数据共享策略。该方法基于混合音素集多语言DNN建模方法,其原理示意图如图5所示。

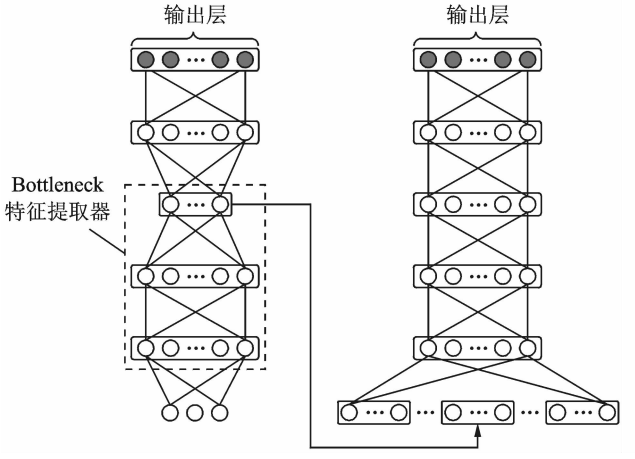

图4　Bottleneck特征提取器示意图

Fig. 4　Illustration of bottleneck feature extractor

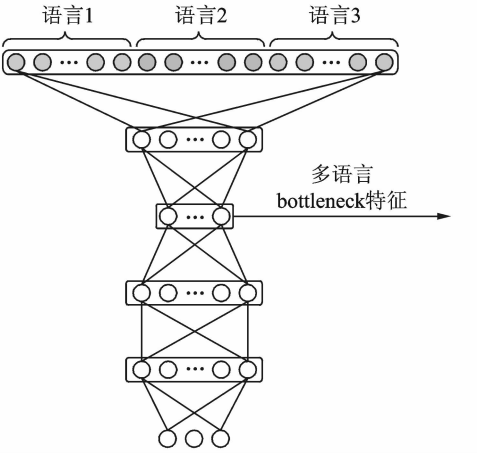

图5　基于多语言bottleneck特征的数据共享示意图

Fig. 5　Illustration of data sharing based on multilingual bottleneck feature

在提取多语言 bottleneck 特征所用的 DNN 模型中，输出层为多语言 HMM 模型的物理状态。其 bottleneck 层具有较少的节点数，并使用线性神经元，以保留尽可能全面的多语言信息。使用多个非目标语言的数据训练得到多语言 bottleneck 特征提取器后，利用该特征提取器对目标语言的训练数据和测试数据提取 bottleneck 特征，并利用目标语言训练数据的 bottleneck 特征来进行声学模型的训练。

Tandem 方法与 hybrid 方法相比，一个优势就是 tandem 方法中用来提取特征的 DNN 模型所用的训练数据并不要求和最终的测试数据非常匹配。因此，在低资源、多语言环境中，基于多语言 bottleneck 特征的数据共享策略有望获得比 hybrid 方法更好的识别效果。

3 低资源语音识别中的声学模型

3.1 子空间高斯混合模型[41]

GMM-HMM 一直是语音识别领域的经典模型，其数学描述简单清晰，参数估计方法完备高效[58]。但是传统的基于 GMM-HMM 的声学建模方法有其自身的局限性，很多情况下都很难仅通过此传统模型架构来达到比较理想的语音识别性能。特别在低语音数据量条件下，这些问题便表现地更加突出，导致系统性能不够理想。这促使广大研究者希望寻找到更加精细的模型结构，从而得到更加强大的模型建模能力，在一定程度上克服传统 GMM-HMM 模型的一些弊端。这其中，子空间高斯混合模型-隐马尔科夫模型(Subspace Gaussian mixture model-hidden Markov model, SGMM-HMM)便是一种很好的声学建模方法[59]。基于子空间思想的 SGMM-HMM 建模方法的每一个隐含马尔可夫模型的状态并不是一个简单的高斯混合模型，而是将音素因子以及说话人因子等引入其中，通过这些因子派生得到一整个子空间模型的参数空间。在传统方法中隐含马尔可夫模型的每一个状态可以用一个独立的高斯混合模型来表示，如状态 j 的高斯混合模型可以表示为

$$p(\boldsymbol{x} \mid j)=\sum_{i=1}^{M_j}\omega_{ji}N(\boldsymbol{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_{ji}) \tag{4}$$

式中：i 表示高斯分量索引号，M_j 表示高斯分量数，N 则表示一个标准的多维高斯正态分布函数。GMM-HMM 模型结构下每一个状态都是独立的，每个状态的模型参数都是独立估计。

SGMM-HMM 模型依旧是基于 HMM 隐含马尔可夫模型，它的每一个状态仍然用一个高斯混合模型来表示，但是它的模型参数的估值方法与传统的方法有所不同，这里给出子空间模型每个状态 j 高斯混合模型的表达式[60]为

$$p(\boldsymbol{x} \mid j)=\sum_{i=1}^{I}\omega_{ji}N(\boldsymbol{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_i) \tag{5}$$

$$\boldsymbol{\mu}_{ji}=\boldsymbol{M}_i \cdot \boldsymbol{v}_j \tag{6}$$

$$\omega_{ji}=\frac{\exp \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j}{\sum_{i'=1}^{I}\exp \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j} \tag{7}$$

式中：$j=\{1,\cdots,J\}$；$i=\{1,\cdots,I\}$，表示 HMM 模型共有 J 个状态，每一状态各由 I 个高斯分量表示，比较典型的 I 值为 $200\leqslant I\leqslant 2\,000$。其中 j 和 i 分别表示状态索引号和高斯分量索引号；$\boldsymbol{x}\in\mathbf{R}^D$，表示一个 D 维的观测向量，也即是一般所指的特征向量；$\boldsymbol{\mu}_{ji}\in\mathbf{R}^D$，表示第 j 个 HMM 状态的第 i 个高斯分量的 D 维均值向量，非直接模型参数，需要计算得到；$\omega_{ji}\in\mathbf{R}$，表示第 j 个 HMM 状态的第 i 个高斯分量的权重，非直接模型参数，需要计算得到；$\boldsymbol{\Sigma}_i\in\mathbf{R}^{D\times D}$，表示第 i 个高斯分量的 $D\times D$ 维的全协方差矩阵，其中所有状态的第 i 个高斯分量的全协方差矩阵都共享，是一个状态间共享模型参数；$\boldsymbol{v}_j\in\mathbf{R}^S$，表示状态 j 的 S 维状态相关向量，其中每一个 HMM 状态都拥有各自的状态相关向量，相互之间不共享；$\boldsymbol{M}_i\in\mathbf{R}^{D\times S}$，表示第 i 个高斯分量的 $D\times S$ 维均值投影矩阵，它是用来将第 j 个状态的状态相关向量 $\boldsymbol{v}_j$ 进行投影，得到第

j 个状态的第 i 个高斯分量的均值向量 $\boldsymbol{\mu}_{ji}$。其中,所有 HMM 状态的第 i 个高斯分量都共享一个均值投影矩阵,是一个状态间共享模型参数;$\boldsymbol{w}_i \in \mathbf{R}^S$,表示第 i 个高斯分量的 S 维权重投影向量,它是用来将第 j 个状态的状态相关向量 $\boldsymbol{v}_j$ 进行投影,得到第 j 个状态的第 i 个高斯分量的权重值。其中,所有 HMM 状态的第 i 个高斯分量都共享一个权重投影向量,是一个状态间共享模型参数;除此之外,还有一个通用背景模型(Universal background model,UBM)并没有在以上公式中列出(后面会再解释),它主要用于子空间模型初始化,声学分数计算时的高斯选择等用途。

子空间高斯混合模型依然沿用 HMM 声学模型架构,用状态输出概率来描述语音特征向量在每个状态的似然度。但它的每一个隐含马尔可夫模型状态除表示状态概率分布外,还将音素因子(或说话人因子等)引入其中,分别建模,与传统 GMM 模型截然不同,主要有如下几点:

(1)子空间高斯混合模型的第 j 个状态的每个高斯分量的均值 $\boldsymbol{\mu}_{ji}$ 和权重 ω_{ji} 已经不像传统 GMM 模型一样,还是直接的模型参数,它们都由与第 j 个状态相关的向量 $\boldsymbol{v}_j$ 经过投影矩阵或者投影向量变换得到。

(2)传统 GMM 模型各个状态的高斯分量相对独立,高斯分量数可以不同,并且高斯数量一般较少(几十个);子空间高斯混合模型各状态的高斯分量相互联系,共享同一个高斯协方差矩阵,并且高斯分量数目相同,并且高斯数量一般很大(几百至上千个)。

(3)传统 HMM 模型各状态的状态信息被包含于所有高斯混合模型参数中,但是子空间模型的每个状态的个性信息仅用状态相关向量 $\boldsymbol{v}_j$ 便可表示。它包含了各状态各自的特殊信息,仅用一个向量便可以将所有状态区分开。

(4)除了所熟知的 GMM 模型参数外,子空间模型还需要一个全局的 UBM,用于模型初始化,分数计算高斯选择等。所有状态都共用一个通用背景模型,它也是状态共享的。

(5)传统 HMM 模型的整个参数空间都是各状态独立的,而子空间模型的整个参数空间由于它独特的表示方式,划分成了状态共享参数空间和状态相关非共享参数空间。如图 6 所示,左边虚线框中用到的 UBM,均值投影矩阵和权重投影向量都属于共享参数空间;而右边虚线框中的状态相关向量 $\boldsymbol{v}_j$ 属于非共享参数空间。

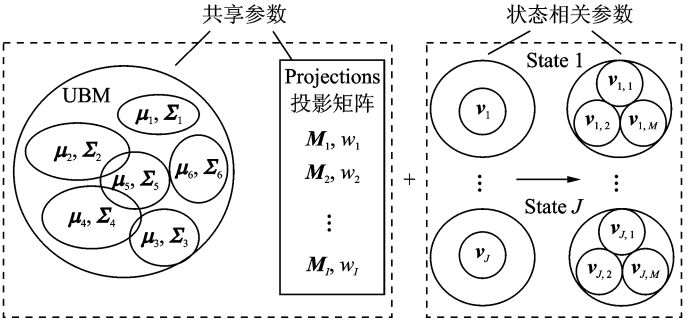

图 6　子空间高斯混合模型参数空间示意图

Fig. 6　Illustration of parameter space of subspace Gaussian mixture model

(6)通常状态相关向量 $\boldsymbol{v}_j$ 的维数 S 与语音特征向量的维数 D 比较相近($S \approx D$)。以一个 $I=16$,$D=39$ 和 $S=40$ 的模型为例,子空间模型与每一个状态相关的参数量只是 $S=40$ 维,而传统的每一个对角方差 GMM 模型的状态相关参数量则为 $I\times(2D+1)=1\ 264$ 维,子空间模型各状态的参数量要远小于传统模型各状态的参数量。所以可以认为子空间的状态参数只位于传统的 GMM 模型参数空间的一个子空间中,这也正是子空间高斯混合模型名称的由来。

3.2 卷积神经网络声学模型[43]

卷积神经网络(Convolutional neural network,CNN)是低资源语音识别中另一种有效的声学建模方法,它可以利用靠近输入层的局部连接网络权重来建模数据的局部特性。典型的 CNN 模型结构如图 7 所示,图中的 CNN 模型使用 fbank 特征[61]作为输入,fbank 特征的不同局部频谱分量在 CNN 的卷积层被分别处理和建模。

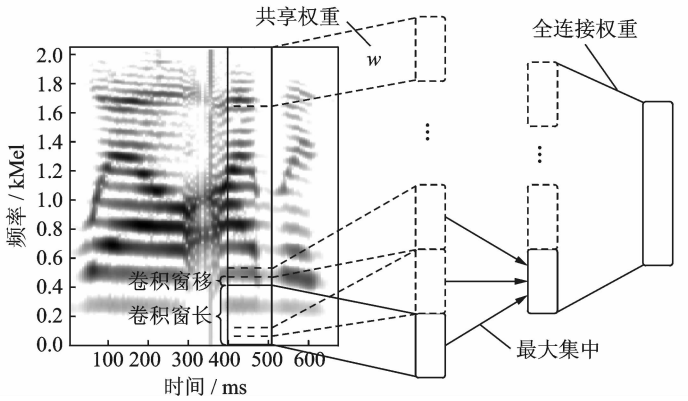

图 7 卷积神经网络应用于语音信号 fbank 特征的示意图

Fig. 7 Illustration of CNN applied to fbank features of speech signals

在基于 CNN 的语音识别声学模型中,输入的特征矢量首先被划分为 N 个互不交叠的频带$\{\boldsymbol{v}_i \mid i=0,\cdots,N-1\}$。譬如,当使用 fbank 特征时,CNN 的输入包含 n 帧连续的特征矢量,可以将其记为

$$\boldsymbol{X}=[\boldsymbol{x}_t,\Delta\boldsymbol{x}_t,\Delta\,\Delta\boldsymbol{x}_t,\boldsymbol{x}_{t+1},\Delta\boldsymbol{x}_{t+1},\Delta\,\Delta\boldsymbol{x}_{t+1},\cdots,\boldsymbol{x}_{t+n},\Delta\boldsymbol{x}_{t+n},\Delta\,\Delta\boldsymbol{x}_{t+n}] \tag{8}$$

式中:$\boldsymbol{x}_t$,$\Delta\boldsymbol{x}_t$ 和 $\Delta\Delta\boldsymbol{x}_t$ 分别代表 fbank 基本特征、fbank 一阶差分特征和 fbank 二阶差分特征。一般情况下,频带数 N 即 fbank 基本特征维数,且频带 $\boldsymbol{v}_i$ 的构成为

$$\boldsymbol{v}_i=[x_t^{(i)},\Delta x_t^{(i)},\Delta\,\Delta x_t^{(i)},x_{t+1}^{(i)},\Delta x_{t+1}^{(i)},\Delta\,\Delta x_{t+1}^{(i)},\cdots,x_{t+n}^{(i)},\Delta x_{t+n}^{(i)},\Delta\,\Delta x_{t+n}^{(i)}]^{\mathrm{T}} \tag{9}$$

式中:$x_t^{(i)}$ 代表特征的第 i 维。将频带$\{\boldsymbol{v}_i \mid i=0,\cdots,N-1\}$中每 s 个相邻的频带划分成一个频带组$\{\boldsymbol{v}_{i+r} \mid r\in 0,\cdots,s-1\}$。该频带组作为 CNN 局部连接网络权重的输入,经过卷积层的权重矩阵、偏置矢量和非线性变换,被用来产生卷积层的输出 $\boldsymbol{h}_i$。当使用全部权重共享[61]策略时,卷积层输出 $\boldsymbol{h}_i$ 的计算方法为

$$\boldsymbol{h}_i=\theta\Big(\sum_{r=0}^{s-1}\boldsymbol{W}_r^{\mathrm{T}}\cdot\boldsymbol{v}_{i+r}+\boldsymbol{b}\Big) \tag{10}$$

式中:θ 为 CNN 卷积层的非线性变换函数,如 sigmoid 函数。$\{\boldsymbol{W}_r \mid r\in 0,\cdots,s-1\}$为 CNN 卷积层作用于频带组$\{\boldsymbol{v}_{i+r} \mid r\in 0,\cdots,s-1\}$的一组权重矩阵,将$\{\boldsymbol{W}_r \mid r\in 0,\cdots,s-1\}$中的矩阵进行拼接,即构成卷积层权重矩阵 $\boldsymbol{W}$。式中的 $\boldsymbol{b}$ 为卷积层的偏置矢量。卷积层权重矩阵 $\boldsymbol{W}$ 与偏置矢量 $\boldsymbol{b}$ 并不随输入频带组编号 i 的变化而变化,从而实现参数共享。在 CNN 卷积层的操作过程中,权重矩阵 $\boldsymbol{W}$ 与偏置矢量 $\boldsymbol{b}$ 每次平移 d 个频带,从而生成一组卷积层频带输出$\{\boldsymbol{h}_j \mid j\in 0,\cdots,M-1\}$。本文将卷积层输入的频带个数 s 称为“卷积窗长”,将卷积层每次平移的频带个数 d 称为“卷积窗移”。通常情况下,卷积窗移满足 $d=1$。

由于实际情况下 CNN 中使用互相交叠的卷积窗,即 $s>d$,因此卷积层的输出可能包含冗余信息。在 CNN 中,最大集中层经常伴随卷积层出现,以达到隐含层节点降维的目的。最大集中操作的数学形式为

$$p_i^m=\max_{j\in i\cdot k,\cdots,(i+1)\cdot k-1}h_j^m \tag{11}$$

式中:p_i^m 为最大集中层的输出节点,i 和 j 分别是最大集中操作前后的频带索引号,m 为相应频带中的节点索引号,k 为最大集中操作输入的频带数目。在 CNN 声学模型中,最大集中层的使用有助于降低不同语音中的频谱差异。例如,不同人说话的音调不同,其语音的共振峰位置在频谱上也会有一定的偏移。当最大集中层作用于若干个频带时,有助于 CNN 模型捕捉到不同说话人的共振峰出现位置,从而降低了不同说话人的差异,使得模型得以鲁棒地估计。但是当最大集中层输入频带数 k 过大时,又会降低模型处理过程中对于频谱的分辨率,从而损伤识别性能。因此 k 的大小控制模型在"区分性"和"不变性"之间寻求平衡[62],有时需要根据具体任务合理调节。经过 CNN 靠近输入层的若干个卷积层和最大集中层变换后,CNN 的隐含层频带被拼接在一起,从而构成前馈全连接结构。经过若干个全连接的隐含层,CNN 最终在输出层使用 softmax 函数实现分类。

在语音识别声学建模任务中,基于 CNN 的模型虽然可以建模语音信号中的局部特性,但是其优化过程中的梯度消失问题和过拟合问题依然是影响模型性能的两个因素。使用 maxout 神经元和 dropout 训练方法来解决基于 CNN 的声学模型优化过程中的梯度消失问题和过拟合问题[63-66]。通过将 maxout 神经元和卷积结构结合,得到一种更适合低资源条件下语音识别声学建模任务的卷积 maxout 神经网络(Convolutional maxout neural network,CMNN)声学模型。

在 CMNN 模型中,CNN 原有的卷积结构、最大集中结构和全连接结构都被保留,而所有非线性函数 θ 由 sigmoid 函数替换为 maxout 函数。CMNN 中的卷积层仍然能够建模语音信号中的局部特性,maxout 神经元的引入使得模型在低资源条件下的优化性能得以提高。图 8 中给出了 CMNN 模型的卷积层中 maxout 操作和最大集中操作的对比关系。在 CMNN 模型的卷积层中,maxout 操作对同一个频带内部相邻位置的神经元节点输出选取最大值。最大集中操作作用于 maxout 操作之后,其输出为不同频带中同一个相对位置处神经元节点的最大值。

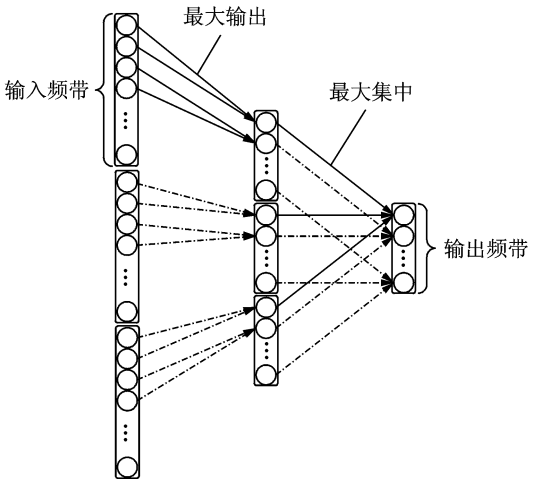

图 8　CMNN 模型卷积层的最大输出操作和最大集中操作对比

Fig. 8　Illustration of maxout and max-pooling applied to convolutional layers of CMNN

相比于前馈全连接的 DNN 基线模型,CMNN 模型具有以下优势:(1)卷积层建模语音信号的帧内先验信息;(2)maxout 神经元解决随机梯度下降(Stochastic gradient descent,SGD)训练过程中的梯度消失问题,使得模型容易优化;(3)dropout 训练方法适合与 maxout 神经元一同使用,防止模型在低资源条件下出现过拟合,增强模型的推广性。

4　低资源语音识别中的神经网络语言模型[42]

在语音识别领域,应用最广泛的语言模型是基于词频统计的 N 元文法模型。理论上模型阶数 N 越高,对下一个词的统计学约束越强,模型的辨识能力也越强,但是 N 元组合在文本库中的出现频次将越低,其概率估计也越不准确。因此在实际系统中,模型的阶数通常不超过 5 阶。

尽管 N 元文法模型性能尚可,具有查询高效等特点,并得到广泛应用,但是仍然存在一些问题,最重要的问题是语言模型的稀疏性问题[67,68],也称"零概率"估值问题。由于自然语言本身具有稀疏性,即满足齐夫定律[69]。自然语言中 20%的单词占据了 80%的文本数据库,即少数的高频词占据了绝大多

数文本数据；相反大量的低频词在文本库中出现次数非常少，甚至没有出现。这种单词分布的不均匀性导致大量的词语组合在训练数据中并没有出现。

采用神经网络对语言模型在连续空间建模可以有效克服传统的词频统计建模方法的不足。神经网络语言模型[70,71]首先定义投影矩阵将离散的词语符号映射到连续空间，形成相应的词矢量特征，某种程度上可以自动形成词义的等价性聚类。实验表明：词法或者语义等相似的词语矢量特征在连续空间表现出比较好的聚集性。将分布式的上文矢量信息输入典型的人工神经网络，并在输出层预测下一个词的概率。

对于神经网络语言模型建模方法，等价类构造函数定义为连续空间的分布式词矢量映射，其在高维空间表现出比较好的聚集性，而相应的预测函数为人工神经网络结构，输出层对应每个词的预测概率。投影矩阵参数和神经网络参数均是基于数据驱动的方法主动从训练数据中学习得到。从等价类构造和预测函数的角度，连续空间建模是离散空间建模的进一步的扩展，同时从人类认知学习的过程，连续空间建模更加符合知识获取过程。和离散空间建模相比，基于神经网络语言模型的连续空间建模在模型性能上取得更好的建模效果。相比于传统的词频建模方法，神经网络语言模型提供了非常好的等价类构造函数与平滑预测函数。

连续空间建模的方法充分挖掘了词矢量特征在“词法、语义”层面的聚集性和神经网络良好的学习能力，对于“零频率”或者“低频率”词语组合的概率估计表现出比较好的平滑效果和泛化能力。事实上对于“零频率”或者“低频率”的词语组合，其在连续空间的近邻并不一定也是“零频率”或者“低频率”，因此通过连续平滑的神经网络函数，其输出概率能够通过词矢量的聚集性得到相对准确的估计。近些年在国内外学者的广泛关注下，神经网络语言模型在自然语言处理以及语音识别领域取得突破性的进展[72-77]。在此基础之上，文献[78～82]研究了在低资源条件下如何有效应对更加严峻的稀疏性问题。

典型的神经网络语言模型包括前馈神经网络语言模型(Feed-forward neural network language model，FNNLM)与递归神经网络语言模型(Recurrent neural network language model，RNNLM)，其结构如图 9 所示。

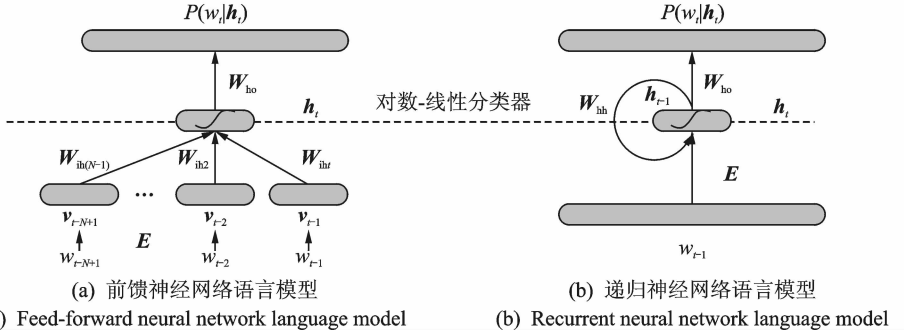

图 9 典型的神经网络语言模型

Fig. 9 Typical neural network language models

前馈神经网络语言模型基于有限历史假设，其输入为上文最近的 $N-1$ 个词；递归神经网络语言模型输入为整个历史上文，突破了有限历史假设，适合长序列建模，其输入为上一个词和上一时刻的隐含层状态。神经网络语言模型的输出层对应字典中每个词的预测概率。

定义词矢量投影矩阵 $\boldsymbol{E}\in R^{|V|\times D}$，其中 $|V|$ 代表字典大小，D 代表词矢量特征的维数；输出预测矩阵为 $\boldsymbol{W}_{ho}\in R^{|V|\times H}$，其中 H 代表隐含层的维数，对于递归神经网络语言模型，其隐含层维数与词矢量特征维数相等，即 $H=D$。对于前馈神经网络语言模型，定义输入层传输矩阵 $\boldsymbol{W}_{ih}\in R^{H\times(N-1)D}=[\boldsymbol{W}_{ih1}^{H\times D}$，

$\boldsymbol{W}_{\text{ih2}}^{H\times D},\cdots,\boldsymbol{W}_{\text{ih}(N-1)}^{H\times D}]$对应 N 阶输入,对于递归神经网络语言模型,定义隐含层递归传输矩阵 $\boldsymbol{W}_{\text{hh}}^{H\times H}$ 对应递归历史状态输入。

给定历史输入 $\boldsymbol{w}_1^{t-1}=[w_1,w_2,\cdots,w_{t-1}]$,预测当前单词 w_t 的概率 $P(w_t|w_1^{t-1})$。首先根据词语在字典中的索引基于投影映射矩阵 $\boldsymbol{E}^{|V|\times D}$ 查表,将离散的词语映射为对应的词矢量输入 $\boldsymbol{v}_1^{t-1}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{t-1}]\in R^{D\times(t-1)}$,然后经过隐含层传输得到隐含层的状态矢量 $\boldsymbol{h}_t\in R^H$。隐含层状态根据不同的模型结构计算如下

$$\begin{aligned}&\text{前馈神经网络语言模型:}\quad \boldsymbol{h}_t=g\Big(\sum_{k=1}^{N-1}\boldsymbol{W}_{\text{in}k}\boldsymbol{v}_{t-k}\Big)\\&\text{递归神经网络语言模型:}\quad \boldsymbol{h}_t=g(\boldsymbol{W}_{\text{hh}}\boldsymbol{h}_{t-1}+\boldsymbol{v}_t)\end{aligned}\tag{12}$$

其中激励函数 $g(\cdot)$通常取 sigmoid(·)或者 tanh(·)函数,即

$$\begin{aligned}\text{sigmoid}(x)&=\frac{1}{1+\mathrm{e}^{-x}}\\\tanh(x)&=\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}\end{aligned}\tag{13}$$

神经网络语言模型的输出层为“对数线性分类器”,又称“最大熵分类器”,通过 Softmax 函数将输出归一化为概率。定义 $\boldsymbol{W}_{\text{ho}}=[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_{|V|}]^{\mathrm{T}}\in R^{|V|\times H}$,其中 $\boldsymbol{\theta}_i\in R^H$ 对应于每个输出层节点。定义句子中第 t 个词在字典中的索引 $q(w_t)=i$,则当前词的概率计算为

$$P(w_t|w_1^{t-1})=P(q(w_t)=i|\boldsymbol{h}_t)=\frac{\exp(\boldsymbol{\theta}_i^{\mathrm{T}}\boldsymbol{h}_t)}{\sum_{j=1}^{|V|}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{h}_t)}\tag{14}$$

给定训练语料文本 T,神经网络语言模型的训练依据交叉熵准则,其交叉熵代价函数为

$$J(\Theta)=-\frac{1}{|T|}\sum_{t=1}^{|T|}\log(P(q(w_t)=i|\boldsymbol{h}_t))=-\frac{1}{|T|}\sum_{t=1}^{|T|}\log\left(\frac{\exp(\boldsymbol{\theta}_i^{\mathrm{T}}\boldsymbol{h}_t)}{\sum_{j=1}^{|V|}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{h}_t)}\right)\tag{15}$$

式中:Θ 为待估参数集合,对于前馈网络语言模型,$\Theta=\{\boldsymbol{E}\in R^{|V|\times D},\boldsymbol{W}_{\text{ih}}\in R^{H\times(N-1)D},\boldsymbol{W}_{\text{ho}}\in R^{|V|\times H}\}$,而对于递归神经网络语言模型,$\Theta=\{\boldsymbol{E}\in R^{|V|\times D},\boldsymbol{W}_{\text{hh}}\in R^{H\times H},\boldsymbol{W}_{\text{ho}}\in R^{|V|\times H}\}$。另外需注意,减小交叉熵代价等效于最大化训练样本的似然度,即最大似然估计。

目标代价函数优化过程通常采用经典的随机梯度下降法进行参数优化。目标函数的梯度为

$$\begin{aligned}\frac{\partial J(\Theta)}{\partial\boldsymbol{\theta}_j}&=-\frac{1}{|T|}\sum_{t=1}^{|T|}\{I(j=q(w_t))-P(q(w_t)=i\mid\boldsymbol{h}_t))\boldsymbol{h}_t\}\\\frac{\partial J(\Theta)}{\partial\boldsymbol{h}_t}&=-\{\theta_{\text{q}(w_t)=i}-\sum\nolimits_{j=1}^{|V|}(P(q(w_t)=i\mid\boldsymbol{h}_t)\boldsymbol{\theta}_j)\}\end{aligned}\tag{16}$$

式中:$I(j=q(w_t))$为指示函数,如果等式成立返回 1,否则返回 0。隐含层状态 $\boldsymbol{h}_t$ 是投影矩阵以及输入矩阵等其他参数的函数,通过复合函数的链式求导法则,可以求得相应参数的梯度参数即 $\nabla\Theta$,采用随机梯度法更新参数如下

$$\Theta_{k+1}=\Theta_k-\eta\cdot\nabla\Theta\tag{17}$$

式中:η 为学习步长,k 代表迭代次数,每次迭代后在开发集上校验模型的性能。如果开发集上性能不再提高或者变差,将终止迭代。通常情况下经过有限次的迭代训练后,训练过程收敛而终止。

5　NIST OpenKWS 关键词评测

自 2013 年起,美国国家标准与技术研究所(National institute of standards and technology,NIST)开始每年举办一次国际关键词检索(OpenKWS)评测。该评测是美国政府大力支持的 IARPA Babel 项目[38]的一部分,受到该项目资助的研究机构必须参加该评测,且评测成绩决定了下一年研究机构所能

获得的资助情况。除了受到 IARPA Babel 项目资助的参赛队伍外，OpenKWS 评测也允许其他参赛队伍参加。NIST OpenKWS 评测的目标是在有限的时间内，利用有限的训练资源建立一套语音识别系统，并进行关键词检索任务。在每次评测中，都会发布一个“意外”语言(Surprise language)，其语种信息事先未知。2013 年的 OpenKWS13 评测语言是越南语，2014 年的 OpenKWS14 评测语言是泰米尔语，2015 年的 OpenKWS15 评测语言为斯瓦希里语，2016 年的 OpenKWS16 评测语言为格鲁吉亚语。

NIST OpenKWS 评测官方指标为最大查询词加权值(Maximum term-weighted value，MTWV)和实际查询词加权值(Actual term-weighted value，ATWV)[83]。由于关键词检索系统的任务是在语音库里找出特定的关键词，包括其出现的起始时间和结束时间。关键词系统可能出现的错误类型有两种：虚警错误(False alarm，FA)和漏报错误。虚警错误是指在没有关键词的位置检出了关键词，而漏报错误是指在有关键词的位置没有检测出关键词。在 NIST OpenKWS 评测中，要求关键词检索系统给每一个可能的关键词赋予一个置信度分数，并使用统一的门限 θ 来判定每个可能的关键词是否为真实的关键词。因此，关键词检索系统的虚警率和漏报率跟系统使用的门限 θ 有关。综合这两种错误类型，衡量关键词检索系统性能的指标 ATWV 定义为

$$\mathrm{ATWV}(\theta)=1-\frac{1}{K}\sum_{w=1}^{K}\left(\frac{\#\,\mathrm{miss}(w,\theta)}{\#\,\mathrm{ref}(w)}+\beta\frac{\#\,\mathrm{fa}(w,\theta)}{T-\#\,\mathrm{ref}(w)}\right) \tag{18}$$

式中：$\#\,\mathrm{ref}(w)$ 为答案中出现的关键词 w 的个数；$\#\,\mathrm{miss}(w,\theta)$ 和 $\#\,\mathrm{fa}(w,\theta)$ 分别为在门限 θ 下关键词 w 的漏报次数和虚警次数；T 代表测试语音库的大小，以 s 为单位，用来作为测试语音中所出现的词个数的估计；K 为不同关键词的个数；β 为一常量，用来对系统的虚警率进行惩罚，在 NIST OpenKWS 评测中，β 的值设置为 999.9。随着门限 θ 的变动，ATWV 的值会有所不同。而在最优的门限值下，系统能够得到 MTWV，这也是评价关键词检索系统性能好坏的一项指标，它代表了一个特定的关键词检索系统能够达到的最优性能。显然，ATWV 和 MTWV 的值越大越好，在理想情况下 ATWV 和 MTWV 的值相等，且两者的最大值均为 1。

图 10 给出了 2015 年 NIST 官方发布的本团队参加评测的 DET(Detection error tradeoff)曲线，当时所构建的系统关键词检索 AWTV 达到 0.539 1，该结果为 NIST OpenKWS15 评测极低资源测试任务公开条件第 1 名[84,85]。

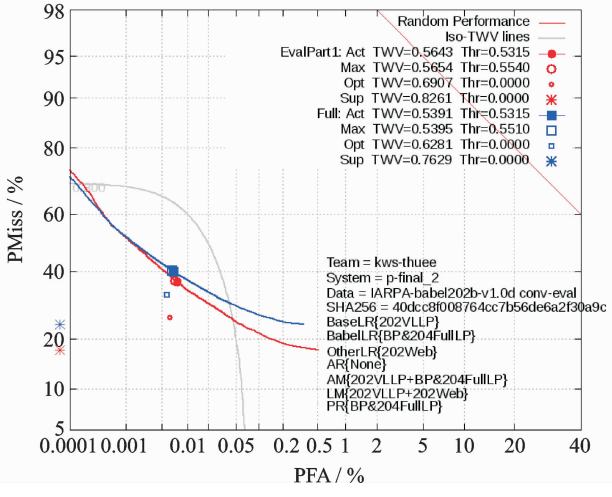

图 10 2015 年 NIST 官方发布的本团队参加 OpenKWS 评测的 DET 曲线

Fig. 10 Official DET curves of our team for OpenKWS evaluation released by NIST

6 结束语

语音识别已经经历了起起落落 60 余年的发展,低资源语音识别也是伴随着语音识别逐步发展起来,令人欣喜的是它已经在一些领域实际应用并取得了不错的效果。近年来,随着深度学习等技术的兴起,低资源语音识别又有了较大的进展。除了本文介绍若干进展之外,另外几个方向也很可能产生新的技术突破:(1)多语言和跨语言声学建模新方法,设计多语言和跨语言的声学建模新方法,利用极少量数据便可对新的语言进行语音识别。(2)端到端的声学建模方法,通过建立语音特征和声学建模基本单元的关系,利用统计方法实现端到端的声学建模,有望使语音识别系统真正摆脱对发音字典的依赖。(3)声学模型和语言模型统一建模,利用统一的模型对声学信息和语言信息建模,更好地挖掘声学和语言之间的关联性与互补性,提高系统性能。

致谢　本文撰写过程中,作者参考了其指导的博士生单煜翔、钱彦旻、史永哲和蔡猛的博士论文的部分文字和图表,在此表示感谢。

参考文献:

[1] Juang B H, Furui S. Automatic recognition and understanding of spoken language—A first step toward natural human-machine communication [J]. Proceedings of the IEEE, 2000, 88(8): 1142-1165.

[2] Deng Li, Huang Xuedong. Challenges in adopting speech recognition [J]. Communications of the ACM, 2004, 47(1):69-75.

[3] Besacier L, Barnard E, Karpov A, et al. Automatic speech recognition for under-resourced languages: A survey [J]. Speech Communication, 2014, 56:85-100.

[4] Thomas S, Ganapathy S, Hermansky H. Multilingual MLP features for low-resource LVCSR systems [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE, 2012:4269-4272.

[5] Vu N T, Imseng D, Povey D, et al. Multilingual deep neural network based acoustic modeling for rapid language adaptation [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 7639-7643.

[6] Chen N F, Ni Chongjia, Chen I-Fan, et al. Low-resource keyword search strategies for Tamil [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015:5366-5370.

[7] Chen N F, Xu H, Xiao X, et al. Li. Exemplar-inspired strategies for low-resource spoken keyword search in Swahili [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 6040-6044.

[8] Laurent A, Fraga-Silva T, Lamel L, et al. Investigating techniques for low resource conversational speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 5975-5979.

[9] Syed A R, Rosenberg A, Kislal E. Supervised and unsupervised active learning for automatic speech recognition of low-resource languages [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 5320-5324.

[10] Muller M, Stuker S, Waibel A. Language adaptive DNNs for improved low resource speech recognition [C]// Interspeech. San Francisco: ISCA, 2016: 3878-3882.

[11] Thomas S, Audhkhasi K, Cui J, et al. Multilingual data selection for low resource speech recognition [C]// Interspeech. San Francisco: ISCA, 2016: 3853-3857.

[12] Zhang Y, Chuangsuwanich E, Glass J, et al. Prediction-adaptation-correction recurrent neural networks for low-resource language speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 5415-5419.

[13] Xie C, Guo W, Hu G, et al. Web data selection based on word embedding for low-resource speech recognition[C]// Interspeech. San Francisco: ISCA, 2016: 1340-1344.

[14] Chen D, Mak B K W. Multitask learning of deep neural networks for low-resource speech recognition [J]. IEEE/ACM Transactions on Speech and Language Processing, 2015(23): 1172-1183.

[15] Mendels G, Cooper E, Soto V, et al. Improving speech recognition and keyword search for low resource languages using web data [C]// Interspeech. Dresden: ISCA, 2015: 829-833.

[16] Cui J, Kingsbury B, Ramabhadran B, et al. Multilingual representations for low resource speech recognition and keyword search [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Scottsdale: IEEE, 2015: 259-

266.

[17] 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]// Interspeech. Dresden: ISCA, 2015: 2132-2136.

[18] Rath S P, Knill K M, Ragni A, et al. Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages [C]// Interspeech. Singapore: ISCA, 2014: 835-839.

[19] Lu L, Ghoshal A, Renals S. Cross-lingual subspace Gaussian mixture models for low-resource speech recognition [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2014(22): 17-27.

[20] 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]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 5592-5596.

[21] Lu L, Ghoshal A, Renals S. Cross-lingual subspace Gaussian mixture models for low-resource speech recognition [J]. IEEE/ACM Transactions on Speech and Language Processing, 2014(22): 17-27.

[22] Zhang W, Fung P. Sparse inverse covariance matrices for low resource speech recognition [J]. IEEE Transactions on Audio Speech and Language Processing, 2013(21): 659-668.

[23] Miao Y, Metze F, Rawat S. Deep maxout networks for low-resource speech recognition [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Olomouc: IEEE, 2013: 398-403.

[24] Miao Y, Metze F, Waibel A. Subspace mixture model for low-resource speech recognition in cross-lingual settings [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2013: 7339-7343.

[25] Xu P, Fung P. Cross-lingual language modeling for low-resource speech recognition [J]. IEEE Transactions on Audio Speech and Language Processing, 2013(21): 1134-1144.

[26] Thomas S, Seltzer M. L, Church K, et al. Deep neural network features and semi-supervised training for low resource speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2013: 6704-6708.

[27] Kanda N, Takeda R, Obuchi Y. Elastic spectral distortion for low resource speech recognition with deep neural networks [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Olomouc: IEEE, 2013: 309-314.

[28] Zhang W, Fung P. Low resource speech recognition with automatically learned sparse inverse covariance matrices [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE. 2012: 4737-4740.

[29] Thomas S, Ganapathy S, Jansen A, et al. Data-driven posterior features for low resource speech recognition applications [C]// Interspeech. Portland: ISCA, 2012: 790-793.

[30] Zhang W, Fung P. Sparse banded precision matrices for low resource speech recognition [C]// Interspeech. Portland: ISCA, 2012: 1912-1915.

[31] 秦楚雄，张连海. 低资源语音识别中融合多流特征的卷积神经网络声学建模方法 [J]. 计算机应用，2016，36(9)：2609-2615.
Qin Chuxiong, Zhang Lianhai. Acoustic modeling approach of multi-stream feature incorporated convolutional neural network for low-resource speech recognition [J]. Journal of Computer Applications, 2016, 36(9): 2609-2615.

[32] 张剑，屈丹，李真. 基于循环神经网络语言模型的 N-best 重打分算法 [J]. 数据采集与处理，2016，31(2)：347-354.
Zhang Jian, Qu Dan, Li Zhen. N-best rescoring algorithm based on recurrent neural network language model [J]. Journal of Data Acquisition and Processing, 2016, 31(2): 347-354.

[33] 张鹏远，计哲，侯炜，等. 一种小资源下语音识别算法设计与优化 [C]//全国人机语音通讯学术会议. 天津：中国中文信息学会，2015.
Zhang Pengyuan, Ji Zhe, Hou Wei, et al. Design and optimization of speech recognition in low resource [C]//National Conference on Man-Machine Speech Communication. Tianjin: CIPSC, 2015.

[34] 伊·达瓦，匂坂芳典，中村哲. 语料资源缺乏的连续语音识别方法的研究 [J]. 自动化学报，2010，36(4)：550-557.
Dawa I, SAGISAKA Yoshinori1, NAKAMURA Satoshi. Investigation of ASR systems for resource-deficient languages [J]. ACTA Automatica Sinica, 2010, 36(4): 550-557.

[35] 刘迪源，郭武. 基于区分性准则的 Bottleneck 特征及其在 LVCSR 中的应用 [J]. 数据采集与处理，2016，31(2)：331-337.
Liu Diyuan, Guo Wu. Discriminative criterion based bottleneck feature and its application in LVCSR [J]. Journal of Data Acquisition and Processing, 2016, 31(2): 331-337.

[36] Elsevier. Engineering village [EB/OL]. http://www.engineeringvillage.com/, 2017-01-09.

[37] Burget L, Goel N, Povey D, et al. Low development cost, high quality speech recognition for new languages and domains [EB/OL]. http://www.clsp.jhu.edu/workshops/archive/ws09/groups/low-development-cost-high-quality-speech-recognition-for-new-languages-and-domains, 2017-01-09.

[38] Harper M. BABEL [EB/OL]. http://www.iarpa.gov/manager_harper.html, 2017-01-09.

[39] Sainath T, Kingsbury B, Metze F, et al. An overview of the base period of the Babel program [N]. SLTC Newsletter, 2013.
[40] 单煜翔. 高效大词汇量连续语音识别解码算法研究与工程化实现 [D]. 北京: 清华大学, 2012.
Shan Yuxiang. Studies and engineering implementation of efficient large vocabulary continuous speech recognition decoding algorithms [D]. Beijing: Tsinghua University, 2012.
[41] 钱彦旻. 低数据资源条件下的语音识别技术新方法研究 [D]. 北京: 清华大学, 2013.
Qian Yanmin. Study on new speech recognition technology under low data resource conditions [D]. Beijing: Tsinghua University, 2013.
[42] 史永哲. 低资源条件神经网络语言模型高效建模与解码方法研究 [D]. 北京: 清华大学, 2014.
Shi Yongzhe. Efficient language modeling and decoding with NNLM for low-resource speech recognition [D]. Beijing: Tsinghua University, 2014.
[43] 蔡猛. 低资源条件下基于深度神经网络的语音识别声学建模研究 [D]. 北京: 清华大学, 2016.
Cai Meng. Research on acoustic modeling based on deep neural networks for low resource speech recognition [D]. Beijing: Tsinghua University, 2016.
[44] Huang X, Acero A, Hon H W. Spoken language processing: A guide to theory, algorithm and system development [M]. Upper Saddle River, New Jersey: Prentice-Hall PTR, 2001.
[45] Siniscalchi S M, Li J, Lee C H. A study on lattice rescoring with knowledge scores for automatic speech recognition [C]// Interspeech. Pittsburgh: ISCA, 2006: 517-520.
[46] Siniscalchi S M, Svendsen T, Lee C H. Toward bottom-up continuous phone recognition [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Kyoto: IEEE, 2007: 566-569.
[47] Siniscalchi S M, Svendsen T, Lee C H. Toward a detector-based universal phone recognizer [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas: IEEE, 2008: 4261-4264.
[48] Cetin O, Kantor A, King S, et al. An articulatory feature-based tandem approach and factored observation modeling [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu: IEEE, 2007: 645-648.
[49] Stuker S, Schultz T, Metze F, et al. Multilingual articulatory features [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hong Kong: IEEE, 2003:144-147.
[50] Qian Yanmin, Xu Ji, Povey Daniel, et al. Strategies for using MLP based features with limited target-language training data [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Honolulu: IEEE, 2011: 354-358.
[51] Qian Yanmin, Liu Jia. Articulatory feature based multilingual MLPs for low-resource speech recognition [C]// Interspeech. Portland: ISCA, 2012: 2602-2605.
[52] Qian Yanmin, Liu Jia. Cross-lingual and ensemble MLPs strategies for low-resource speech recognition [C]// Interspeech. Portland: ISCA, 2012: 2582-2585.
[53] Yan Zhijie, Huo Qiang, Xu Jian. A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR [C]// Interspeech. Lyon: ISCA, 2013: 104-108.
[54] Hermansky H, Ellis D P W, Sharma S. Tandem connectionist feature extraction for conventional HMM systems [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Istanbul: IEEE, 2000:1635-1638.
[55] Grezl F, Karafiat M, Kontar S, et al. Probabilistic and bottle-neck features for LVCSR of meetings [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu: IEEE, 2007: 757-760.
[56] Sainath T N, Kingsbury B, Sindhwani V, et al. Low-rank matrix factorization for deep neural network training with high-dimensional output targets [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2013: 6655-6659.
[57] Xue Jian, Li Jinyu, Gong Yifan. Restructuring of deep neural network acoustic models with singular value decomposition [C]// Interspeech. Lyon: ISCA, 2013: 2365-2369.
[58] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2):257-286.
[59] Povey D, Burget L, Agarwal M, et al. Subspace Gaussian mixture models for speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas: IEEE, 2010: 4330-4333.
[60] Povey D, Burget L, Agarwal M, et al. The subspace Gaussian mixture model—A structured model for speech recognition. Computer Speech and Language [J]. 2011, 25(2): 404-439.
[61] Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2013: 8614-8618.
[62] Deng Li, Abdel-Hamid O, Yu Dong, et al. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion [C]// International Conference on Acoustics, Speech and Signal Processing (IC-

ASSP). Vancouver: IEEE, 2013: 6669-6673.
[63] Cai Meng, Liu Jia. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition [J]. Speech Communication, 2016(77): 53-64.
[64] Cai Meng, Shi Yongzhe, Kang Jian, et al. Convolutional maxout neural networks for low-resource speech recognition [C]// International Symposium on Chinese Spoken Language Processing (ISCSLP). Singapore: IEEE, 2014: 133-137.
[65] Cai Meng, Shi Yongzhe, Liu Jia. Stochastic pooling maxout networks for low-resource speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 3266-3270.
[66] Cai Meng, Shi Yongzhe, Liu Jia. Deep maxout neural networks for speech recognition [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Olomouc: IEEE, 2013: 291-296.
[67] Katz S M. Estimation of probabilities from sparse data for the language model component of a speech recognizer [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1987(3): 400-411.
[68] Kneser R, Ney H. Improved backing-off for n-gram language modeling [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Detroit: IEEE, 1995: 181-184.
[69] Wikipedia. 齐夫定律(Zipf-law)[EB/OL]. http://en.wikipedia.org/wiki/Zipfs_law, 2017-01-09.
[70] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003(3): 1137-1155.
[71] Tomas M, Martin K, Lukas B, et al. Recurrent neural network based language model [C]// Interspeech. Makuhari: ISCA, 2010: 1045-1048.
[72] Stefan K, Tomas M, Martin K, et al. Recurrent neural network based language modeling in meeting recognition [C]// Interspeech. Florence: ISCA, 2011: 2877-2880.
[73] Tomas M, Anoop D, Stefan K, et al. Empirical evaluation and combination of advanced language modeling techniques [C]// Interspeech. Florence: ISCA, 2011: 605-608.
[74] So L H, Allauzen R, Wisniewski G, et al. Training continuous space language models: Some practical issues [C]// Conference on Empirical Methods in Natural Language Processing (EMNLP). MIT: ACL, 2010: 778-788.
[75] Schwenk H, Gauvain J L. Connectionist language modeling for large vocabulary continuous speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Orlando: IEEE, 2002: 765-768.
[76] Schwenk H, Gauvain J L. Using neural network language models for LVCSR [C]// 2004 Rich Transcriptions Workshop. Pallisades: NIST, 2004.
[77] Schwenk H, Dchelotte D, Gauvain J L. Continuous space language models for statistical machine translation [C]// International Conference on Computional Linguistics (COLING). Sydney: ACL, 2006: 723-730.
[78] Shi Yongzhe, Zhang Weiqiang, Cai Meng, et al. Efficient one-pass decoding with NNLM for speech recognition [J]. IEEE Signal Processing Letters, 2014(21): 377-381.
[79] Shi Yongzhe, Zhang Weiqiang, Cai Meng, et al. Empirically combining unnormalized NNLM and back-off N-gram for fast N-best rescoring in speech recognition [J]. EURASIP Journal on Audio, Speech, and Music Processing, 2014, 2014: 19.
[80] Shi Yongzhe, Zhang Weiqiang, Liu Jia, et al. RNN language model with word clustering and class-based output layer [J]. EURASIP Journal on Audio, Speech, and Music Processing, 2013, 2013: 22.
[81] Shi Yongzhe, Zhang Weiqiang, Cai Meng, et al. Variance regularization of RNNLM for speech recognition [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 4931-4935.
[82] Shi Yongzhe, Zhang Weiqiang, Cai Meng, et al. Temporal kernel neural network language modeling[C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2013: 8247-8251.
[83] NIST. KWS15 keyword search evaluation plan [EB/OL]. http://www.nist.gov/itl/iad/mig/upload/KWS15-evalplan-v05.pdf. 2017-01-09.
[84] Cai Meng, Lü Zhiqiang, Song Beili, et al. The THUEE system for the OpenKWS14 keyword search evaluation [C]// International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane: IEEE, 2015: 4734-4738.
[85] Zhang Zhuo, Zhang Weiqiang, Shen Kaixiang, et al. THUEE language modeling method for the OpenKWS 2015 evaluation [C]// International Symposium on Signal Processing and Information Technology (ISSPIT). Abu Dhabi: IEEE, 2015: 507-511.

作者简介：

刘加(1954-)，男，教授，博士生导师，语音与音频信号处理，E-mail: liuj@tsinghua.edu.cn。

张卫强(1979-)，男，副研究员，语音与音频信号处理，E-mail: wqzhang@tsinghua.edu.cn。