文章编号:2095-6134(2015)01-097-06

低数据资源条件下基于 Bottleneck 特征与 SGMM 模型的语音识别系统*

吴蔚澜[1,2], 蔡 猛[3], 田 垚[3], 杨晓昊[3], 陈振锋[1,2], 刘 加[3], 夏善红[2†]

(1 中国科学院大学, 北京 100190; 2 中国科学院电子学研究所 传感技术国家重点实验室, 北京 100190;
3 清华大学电子工程系 清华信息科学与技术国家实验室, 北京 100084)

(2014 年 2 月 27 日收稿; 2014 年 3 月 7 日收修改稿)

Wu W L, Cai M, Tian Y, 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.

摘 要 语音识别系统需要大量有标注训练数据,在低数据资源条件下的识别性能往往不理想.针对数据匮乏问题,本文先研究子空间高斯混合声学模型通过参数共享减少待估计的参数规模,并使用基于最大互信息准则的区分型训练技术提高识别精度;而后在特征层面应用基于深度神经网络的 Bottleneck 特征来达到特征提取和降维的目的;最后将上述研究成果结合并构建了低资源条件下的语音识别系统.在国际标准的 OpenKWS 2013 数据库上的实验结果表明,本文的技术能够有效改善低资源条件下的系统识别性能,相比基线系统有 12% 左右的词错误率降低.

关键词 语音识别; 低资源; 声学模型; 声学特征

中图分类号:TP391.42 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2015.01.016

Bottleneck features and subspace Gaussian mixture models for low-resource speech recognition

WU Weilan[1,2], CAI Meng[3], TIAN Yao[3], YANG Xiaohao[3], CHEN Zhenfeng[1,2], LIU Jia[3], XIA Shanhong[2]

(1 *University of Chinese Academy of Sciences, Beijing* 100190, *China*; 2 *State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing* 100190, *China*; 3 *Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract State-of-the-art speech recognition systems often depend on a lot of training data, but perform poorly when limited data is available. In this paper, we study speech recognition systems under low-resource condition. The subspace Gaussian mixture (SGMM) model is first applied to reduce the number of parameters. The model is further enhanced by discriminative training based on maximum mutual information criterion. The bottleneck features based on deep neural networks are then studied to make robust feature extraction. The SGMM model and the bottleneck features are finally combined to produce a novel speech recognition system under low-resource condition. On the

* 国家自然科学基金(61005019,61273268,61370034,90920302)和北京市自然科学基金(KZ201110005005)资助

† 通信作者,E-mail:shxia@mail.ie.ac.cn

standard OpenKWS 2013 evaluation corpus, experimental results show the combination of the two technologies brings substantial relative improvement of about 12% over the baseline system.

Key words speech recognition; low-resource; acoustic model; acoustic feature

低资源条件下的语音识别问题是指,进行语音识别时,可获取的语音语料等数据资源匮乏,该问题是语音识别领域的一个难点[1-2]. 传统的语音识别技术有 3 大基石,即:隐马尔科夫模型(hidden Markov model, HMM)[3]、高斯混合模型(Gaussian mixture model, GMM)和梅尔频标倒谱系数(Mel frequency cepstrum coefficient, MFCC)[4]等短时频谱特征. 基于 GMM-HMM 的声学模型依赖其简单清晰的数学表示与完备的参数估计方法,一直以来都是语音识别声学建模领域的常青树[3]. 但是 GMM-HMM 声学模型并非完美,特别是在低资源条件下有如下局限性.

GMM-HMM 模型在拥有无限大的训练数据时可以在理论上获得最优性能,该条件在现实中不可能满足. 且不论训练数据能过千小时的常用语言诸如汉语、英语,其距离无限多也存在相当大的差距. 若是低资源条件下的小语种如越南语、老挝语,其可采集的库只有几十小时甚至几小时,此时更与理论最优相去甚远.

GMM-HMM 模型参数规模庞大,在低资源条件下无法获得稳健的参数估值. 中等规模的 GMM-HMM 声学模型可以拥有百万量级的参数. 这种情况下只有拥有大量的学习样本方能获得鲁棒而准确的模型参数估值. 在低资源条件下,这会引发数据稀疏问题,从而导致 GMM-HMM 模型参数估值的不准确.

GMM-HMM 模型各状态间相互独立,参数也互不共享. 虽然依照传统的建模方法可以通过建模过程中的决策树共享获得聚类的声学状态,但是状态一经确定相互之间便完全独立,参数也互不共享完全分开. 这给低资源条件下稳健的模型参数估计带来许多麻烦.

综上所述,虽然 GMM-HMM 声学模型在语音识别领域取得了很好的成绩,但是由于其自身的种种局限性,在低资源条件下很难通过该模型获得理想的语音识别性能. 为了克服传统 GMM-HMM 模型的缺点,众多专家在近年来不断地探寻新的声学建模方法,在这个过程中出现的基于子空间高斯混合模型(subspace Gaussian mixture model, SGMM)的隐马尔科夫声学模型[5-6]和基于深度神经网络(deep neural network, DNN)的隐马尔科夫声学模型[7-8]都是具有划时代意义的研究成果.

本文针对低资源条件下语音识别系统性能差的问题,对声学模型和声学特征展开研究. 声学模型层面将 SGMM 模型应用于资源匮乏的语音识别系统,并对其进行基于最大互信息准则(maximum mutual information, MMI)的区分性训练[9-10]. 声学特征层面将传统声学特征替换为基于 DNN 的 Bottleneck 声学特征[11]. 实验结果表明,采用以上方法,低资源条件下的语音识别系统性能获得了极大提升.

1 低资源条件下基于 GMM-HMM 的语音识别基线系统

1.1 GMM-HMM 声学模型

HMM 通过一个双重的随机过程来模拟人类发音时产生的语音流,它的组成包括可观测的单状态输出和隐含层状态. 第一重随机过程是一个隐马尔可夫随机链,它包含一串外部不可观测的马尔科夫隐含状态,对应真实语音流的内容各状态按照一定的转移概率进行跳转. 可观测的单状态输出由第二重随机过程表示,它描述语音信号的短时频谱特征.

HMM 的状态输出概率可用多种概率分布函数进行建模,这里采用的 GMM-HMM 就是将 HMM 的状态表示为独立的多高斯混合模型,该模型数学表示公式如下

$$p(o_t \mid s_j) = \sum_{m=1}^{M} w_m \cdot \frac{1}{\sqrt{(2\pi)^D \mid \boldsymbol{\Sigma}_m \mid}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_m)^{\mathrm{T}} \right], \quad (1)$$

式(1)是状态 s_j 对观测序列 $\boldsymbol{o}_t$ 的输出概率表示,其中第 m 个高斯混合分量的均值、方差与权重分别用 $\boldsymbol{\mu}_m$、$\boldsymbol{\Sigma}_m$、w_m 表示.

1.2 实验配置

实验对象为低资源条件下的越南语语音识别系统.

实验数据来自美国情报高等计划署 IARPA 组织的 BABEL 计划,具体数据资料如下[12]:

- 实验语言为越南语
- 训练集包含 1 042 个文件拥有 80 h 语音数据
- 开发集包含 132 个文件拥有 10 h 语音数据

实验采用的评价为词错误率(word error rate, WER),该指标可以细分为插入错误、替代错误和删除错误,具体计算公式[13]如下:

$$\text{替代错误率} = \frac{\text{替代词数}}{\text{实际总词数}} \times 100\%,$$

$$\text{插入错误率} = \frac{\text{插入词数}}{\text{实际总词数}} \times 100\%,$$

$$\text{删除错误率} = \frac{\text{删除词数}}{\text{实际总词数}} \times 100\%,$$

$$\text{词正确率} = \frac{\text{正确词数}}{\text{实际总词数}} \times 100\%,$$

词错误率(WER) = 替代错误率 + 插入错误率 + 删除错误率. (2)

针对该评价指标,0% 为最佳得分,最差得分可以超过 100%. 本文实验环境的搭建基于 kaldi 工具包[14].

1.3　低资源条件下的基线系统

低资源条件下的语音识别基线系统描述如下:

- 声学特征层面,提取感知线性预测(perceptual linear predictive, PLP)声学特征,拼接上 1、2 阶差分,取前后各 4 帧进行拼接,将获取的特征矢量经过线性判别分析(linear discriminant analysis, LDA)降维,而后经过最大似然线性变换(maximum likelihood linear transformation, MLLT).
- 声学模型层面,采用说话人自适应训练(speaker adaptive training, SAT)后的 GMM-HMM 声学模型.
- 采用 3-GRAM 语言模型,训练数据来自 BABEL 计划的越南语数据库. 发音字典包含 6 200个单词. 音素集基于 BABEL 计划提供的越南语字典,决策树采用数据驱动技术,无先验语言知识要求.

此外,本文中,无论基线系统或基于后续技术的改进系统,都保持语言模型,发音字典以及音素集合不变.

调整模型参数获取低资源条件下最优的越南语语音识别基线系统,将其简写为 PLP-GMM,见表 1.

表 1　低资源越南语语音识别基线系统性能

Table 1　Performance of baseline Vietnam speech recognition under low-resource condition

系统描述	词错误率(开发集)/%
PLP-GMM	66.2

表 1 中基线系统性能达到国际最新报道相同实验配置下的相近水平[12]. 由表 1 可知,低资源条件下传统声学模型与声学特征表现较差,本文将主要从这两方面展开研究,以期改善低资源语音识别系统的性能.

2　基于 SGMM-HMM 的声学建模研究

2.1　SGMM-HMM 声学模型

SGMM-HMM 是基于子空间思想的声学建模方法,其引入音素因子与说话人因子,虽然每一状态的声学似然与 GMM-HMM 一样,为多高斯分量的叠加,但是模型参数与具体计算方法发生了极大的变化,其最简单的模型可以表示为如下公式:

$$p(\boldsymbol{x} \mid j) = \sum_{i=1}^{I} \omega_{ji} N(\boldsymbol{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i), \quad (3)$$

$$\boldsymbol{\mu}_{ji} = \boldsymbol{M}_i \cdot \boldsymbol{v}_j, \quad (4)$$

$$\omega_{ji} = \frac{\exp \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{v}_j}{\sum_{i'=1}^{I} \exp \boldsymbol{w}_{i'}^{\mathrm{T}} \boldsymbol{v}_j}. \quad (5)$$

式(2)至式(5)中,$j \in \{1, \cdots, J\}$ 表示 HMM 模型状态数;$i \in \{1, \cdots, I\}$ 表示每一状态有多少高斯分量叠加;$\boldsymbol{x} \in R^D$ 为 D 维的特征向量;$\boldsymbol{\Sigma}_i \in R^{D \times D}$ 为状态之间相互共享的全协方差矩阵;$\omega_{ji} \in R$ 为混合权重;$\boldsymbol{\mu}_{ji} \in R^D$ 为 D 维均值向量;$\boldsymbol{v}_j \in R^S$ 为状态间互不共享的状态相关量;$\boldsymbol{M}_i \in R^{D \times S}$ 是均值投影矩阵;$\boldsymbol{w}_i \in R^S$ 是对应投影权重. 除上述公式外还需要一个通用背景模型(universal background model, UBM)用于 SGMM 模型初始化等用途.

GMM 模型的参数空间中各状态相互独立,SGMM 模型的参数空间可以划分为状态相关不共享与状态共享两部分. 状态相关不共享部分包含 $\boldsymbol{v}_j$,状态共享部分包含 UBM、$\boldsymbol{M}_i$、$\boldsymbol{w}_i$. 正是由于

以上特性使得 SGMM 模型更加紧凑灵活. 式(2)至式(5)是对 SGMM 模型的基本描述,为使其拥有更精细的建模能力,我们引入子状态(sub-state)的概念,每一个状态 j 拥有 M_j 个子状态,索引为 $1 \leqslant m \leqslant M_j$,每个子状态拥有与自身相关的向量 $\boldsymbol{v}_{jm}$ 与加权因子 c_{jm}. 由此对 SGMM 模型的数学描述可以扩展为:

$$p(\boldsymbol{x} \mid j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} \omega_{jmi} N(\boldsymbol{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i), \quad (6)$$

$$\boldsymbol{\mu}_{jmi} = \boldsymbol{M}_i \cdot \boldsymbol{v}_{jm}, \quad (7)$$

$$\omega_{jmi} = \frac{\exp \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{v}_{jm}}{\sum_{i'=1}^{I} \exp \boldsymbol{w}_{i'}^{\mathrm{T}} \boldsymbol{v}_{jm}}. \quad (8)$$

在式(7)的基础上引入说话人相关状态向量 $\boldsymbol{v}^{(s)}$ 得到下式,将说话人相关信息与无关信息分别建模,达到对语音信号更准确的描述.

$$\boldsymbol{\mu}_{jmi}^{(s)} = \boldsymbol{M}_i \cdot \boldsymbol{v}_{jm} + N_i \boldsymbol{v}^{(s)}. \quad (9)$$

2.2 SGMM 与 GMM 参数规模比较

GMM 模型参数规模见表 2[15],这里假设特征维数为 D,GMM-HMM 模型拥有状态数为 J,每个状态由 I 个高斯分量叠加而成.

表 2 GMM-HMM 模型参数规模统计[15]

Table 2 Parameters of GMM-HMM model

参数名称	参数规模
高斯权重	$D \times J \times I$
高斯均值	$D \times J \times I$
对角协方差矩阵	$J \times I$

SGMM 模型参数规模见表 3[15],这里假设状态相关向量空间为 S 维,每一状态含有 M 个子状态.

表 3 SGMM-HMM 模型参数规模统计[15]

Table 3 Parameters of SGMM-HMM model

参数名称	参数规模
高斯权重投影向量	$D \times S \times I$
高斯均值投影矩阵	$I \times D \times (D+1)/2$
全协方差矩阵	$S \times I$
子状态权重	$S \times J \times M$
子状态相关向量	$J \times M$
UBM 高斯权重	$I \times D \times (D+1)/2$
UBM 高斯均值	$I \times D$
UBM 高斯方差	I

分析表 2 与表 3 中 GMM 模型与 SGMM 模型的参数规模,可知,相同训练数据情况下 SGMM 模型参数总量更少,在低资源条件下 SGMM 模型可以较好地解决数据稀疏问题,从而获得更鲁棒更准确的参数估值;GMM 模型各状态之间的独立性导致低资源情况下很多状态无法获取上佳的模型估值,而 SGMM 拥有参数共享空间,可以用到几乎所有训练数据,各状态即使在数据资源匮乏的条件下也能拥有稳健的估值;SGMM 模型可以灵活地根据训练数据调整子状态数,在对总参数规模几乎无影响的情况下更精细地对各状态进行建模.

2.3 基于 SGMM 模型的低资源语音识别系统

在低数据资源条件下,采用 SGMM-HMM 声学模型,同时对该模型进行基于 MMI 准则的区分性训练,声学特征和系统其他层面与基线系统保持一致. 将这 2 套系统分别简写为 PLP-SGMM 与 PLP-SGMM-MMI,调整它们的模型参数以获取低资源条件下最优的越南语语音识别性能,并将其与基线系统进行比较,比较结果见表 4.

表 4 基于 SGMM 的低资源语音识别系统性能

Table 4 Performance of SGMM system on Vietnam speech recognition under low-resource condition

系统描述	词错误率(开发集)/%
PLP-GMM	66.2
PLP-SGMM	62.8
PLP-SGMM-MMI	59.6

由表 4 看出,低资源条件下 SGMM 模型远胜 GMM 模型,将其基于 MMI 准则进行区分性训练后,系统的识别效果得到了更大幅度的提升. 由此可知基于 SGMM 的声学建模方法可以有效地改善低资源条件下的语音识别性能.

3 基于 DNN-HMM 的 Bottleneck 特征研究

3.1 DNN-HMM 声学模型

DNN 模型是人工神经网络(artificial neural network,ANN)的扩展. 相比于传统的 ANN 模型它具有更深的隐含层数,更强大的模型刻画能力[16].

DNN 是一个包含输入层、隐含层和输出层的前馈神经网络. 隐含层节点的线性变换基于 logistic 函数,描述参见下式:

$$y_j = \mathrm{logistic}(x_j),\ x_j = b_j + \sum_i y_i w_{ij}, \quad (10)$$

式中,x_j 为隐含层节点的上层输入总和,y_i 为上层输出,b_j 为 x_j 偏移量,w_{ij} 为权重.

DNN 输出层采用 softmax 函数:

$$p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)}, \quad (11)$$

式(11)中输出类别的索引由 k 表示.

DNN 输出层采用 softmax 函数时,其代价函数用交叉熵来表示,定义为

$$C = -\sum_j d_j \log p_j, \quad (12)$$

式中,d_j 表示目标概率,当其为训练数据所属类别时等于 1,其他情况等于 0,p_j 表示 DNN 输出概率.

本文中 DNN 的训练方案为预训练与有监督训练相结合. 预训练采用受限波尔兹曼机(restricted Boltzmann machine, RBM)方法[17-18].

有监督训练采用"错误反向传播"算法,具体描述参见下式:

$$(\boldsymbol{W}_{n+1}^l, \boldsymbol{b}_{n+1}^l) = (\boldsymbol{W}_n^l, \boldsymbol{b}_n^l) + \Delta(\boldsymbol{W}_n^l, \boldsymbol{b}_n^l). \quad (13)$$

该算法对最优参数方向的找寻是基于一阶梯度的. 式(13)中,DNN 的隐含层数为 L,进行 n 次迭代,l 层第 n 次迭代的偏移量与权重分别表示为 $\boldsymbol{b}_n^l$ 和 $\boldsymbol{W}_n^l$,第 n 次迭代的偏移向量与权重矩阵变化量用 $\Delta(\boldsymbol{W}_n^l, \boldsymbol{b}_n^l)$ 表示,计算公式如下:

$$\Delta(\boldsymbol{W}_n^l, \boldsymbol{b}_n^l) = \boldsymbol{\alpha}\Delta(\boldsymbol{W}_{n-1}^l, \boldsymbol{b}_{n-1}^l) - \boldsymbol{\varepsilon}\frac{\partial \boldsymbol{C}}{\partial(\boldsymbol{W}_n^l, \boldsymbol{b}_n^l)}, \quad (14)$$

式中,$\boldsymbol{\alpha}$ 为平滑一阶梯度的"动量"参数,$\boldsymbol{\varepsilon}$ 与 $\boldsymbol{C}$ 分别表示学习速率和代价函数.

DNN 的声学建模方法是用神经网络的输出层来代替 GMM 对 HMM 的状态概率进行建模. 公式为

$$\boldsymbol{P}_{o|s}(\boldsymbol{o}\mid\boldsymbol{s}) = \frac{\boldsymbol{P}_{s|o}(\boldsymbol{s}\mid\boldsymbol{o})}{\boldsymbol{P}_s(\boldsymbol{s})} \cdot \mathrm{const}(\boldsymbol{o}), \quad (15)$$

式中,$\boldsymbol{s}$ 是 HMM 的绑定 3 音素状态,$\boldsymbol{o}$ 是观测向量亦称为测试样本,$\boldsymbol{P}_{o|s}(\boldsymbol{o}\mid\boldsymbol{s})$ 是 HMM 状态输出概率,DNN 输出是 $\boldsymbol{P}_{s|o}(\boldsymbol{s}\mid\boldsymbol{o})$,由训练数据估计的 3 音素状态先验概率是 $\boldsymbol{P}_s(s)$,实际应用中可以忽略公式中的常数 $\mathrm{const}(\boldsymbol{s})$.

3.2　基于 DNN 模型的 Bottleneck 声学特征

传统的 Bottleneck 声学特征由多层感知器(multi-layer perceptron,MLP)获取,它是一种非线性的特征变换与降维技术[19-20],本文基于 DNN 模型来生成 Bottleneck 声学特征. 将基于 LDA-MLLT 与特征最大似然线性回归(feature maximum likelihood linear regression,fMLLR)技术的 PLP 特征取前后各 4 帧进行拼接,应用 LDA 降维,送入提前训练好的 DNN 模型中,该模型在 Bottleneck 层之前拥有 5 个隐含层,在其之后拥有 1 个隐含层,Bottleneck 层节点数为 42,其他层为 1 024,输入层节点数为 250,输出层节点数为 4 514,从 DNN 中的 Bottleneck 层提取出 Bottleneck 特征,将其与之前的特征进行拼接,再次 LDA 降维,应用于低资源条件下基于 SGMM-HMM 与 SGMM-HMM + MMI 技术的越南语语音识别系统.

3.3　基于 Bottleneck 特征的低资源语音识别系统

在低数据资源条件下,将基于 DNN 模型的 Bottleneck 特征应用于 SGMM 与 SGMM + MMI 系统,系统除声学特征与声学模型外其他与基线系统保持一致,将这 2 套系统简写为 BN-SGMM 与 BN-SGMM-MMI 并将它们与之前的系统进行比较,结果见表 5.

表 5　基于 DNN 的低资源语音识别系统性能

Table 5　Performance of Vietnam speech recognition based on DNN under low-resource condition

系统描述	词错误率(开发集)/%
PLP-GMM	66.2
PLP-SGMM	62.8
PLP-SGMM-MMI	59.6
BN-SGMM	60.0
BN-SGMM-MMI	57.7

从表 5 看出,低资源条件下结合 Bottleneck 声学特征的 SGMM 与 SGMM + MMI 系统在之前基础上性能获得极大提升. 最优的 BN-SGMM-MMI 系统性能比基线水平提升 12%. SGMM 系统的解码速度约为 GMM 系统的 2 至 3 倍,将声学特征换为 Bottleneck 特征后解码速度可提升 25% 左右,可见基于 DNN 模型的 Bottleneck 声学特征与 SGMM 声学模型能够有效地解决低资源条件下语音识别性能差的问题.

4 结论

本文针对低资源条件下语音识别系统性能差的问题展开研究. 声学模型层面用 SGMM 模型代替传统的 GMM 模型,并对其进行基于 MMI 准则的区分性训练;声学特征层面用基于 DNN 声学模型的 Bottleneck 特征代替传统特征. 最后将 2 项技术同时应用于低资源条件下的语音识别系统,使得系统性能较基线水平有 12% 的提升,有效地改善了低资源条件下语音识别系统的识别性能.

参考文献

[1] Cui X, Xue J, Dognin P L, et al. Acoustic modeling with bootstrap and restructuring for low-resourced languages[C]// Interspeech. 2010: 2 974-2 977.

[2] Vu N T, Schlippe T, Kraus F, et al. Rapid bootstrapping of five eastern european languages using the rapid language adaptation toolkit[C]// Interspeech. 2010: 865-868.

[3] Rabiner L R. A Tutorial on hidden markov models and selected applications in speech recognition[J]. Proceedings of IEEE, 1989, 77(2):257-286.

[4] Davis S, Mermelstein P. Comparison of parametric representations formonosyllable word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357-366.

[5] Povey D, Burget L, Agarwal M, et al. Subspace Gaussian mixture models for speech recognition[C]// Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010: 4 330-4 333.

[6] Povey D, Burget L, Agarwal M, et al. The subspace Gaussian mixture model: a structured model for speech recognition[J]. Computer Speech and Language, 2011, 25(2):404-439.

[7] Dahl G, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition [J]. IEEE Trans on Audio, Speech and Language Processing, 2012, 20(1): 30-42.

[8] Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks [C] // Interspeech. 2011: 437-440.

[9] Normandin Y. Hidden Markov models, maximum mutual information estimation, and the speech recognition problem [D]. Canada: McGill University, 1991.

[10] He X D, Deng L, Chou W. Discriminative learning in sequential pattern recognition[J]. IEEE Signal Processing Magazine, 2008, 14(1):14-36.

[11] Yu D, Seltzer M L. Improved bottleneck features using pretrained deep neural networks [C] // INTERSPEECH. 2011: 237-240.

[12] IARPA. OpenKWS13 keyword search evaluation [EB/OL]. (2013-01-25)[2014-05-15]. http: // www. nist. gov/itl/iad/mig/upload/OpenKWS13.

[13] 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.

[14] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]// Proc ASRU. 2011: 1-4.

[15] 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.

[16] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE, Signal Processing Magazine, 2012, 29(6): 82-97.

[17] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1 527-1 554.

[18] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504-507.

[19] Fontaine V, Ris C, Boite J M. Nonlinear discriminant analysis for improved speech recognition [C] // Eurospeech. 1997.

[20] Grézl F, Karafiát M, Kontár S, et al. Probabilistic and bottle-neck features for LVCSR of meetings [C] // Proc ICASSP. 2007(4):757-761.