北京大学学报(自然科学版) 第 54 卷 第 2 期 2018 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 2 (Mar. 2018)
doi: 10.13209/j.0479-8023.2017.154

基于瓶颈特征的藏语拉萨话连续语音识别研究

周楠 赵悦† 李要嫱 徐晓娜 才旺拉姆 吴立成

中央民族大学信息工程学院, 北京 100081; † 通信作者, E-mail: zhaoyueso@muc.edu.cn

摘要 基于从深度神经网络提取的瓶颈特征具有语音长时相关性和紧凑表示的特点, 将瓶颈特征及其与MFCC的复合特征用于藏语连续语音识别任务中, 可以代替传统的MFCC特征进行GMM-HMM声学建模。在藏语拉萨话连续语音识别任务中的实验表明, 瓶颈特征的复合特征取得比深度神经网络后验特征和单瓶颈特征更好的识别表现。

关键词 藏语拉萨话; 连续语音识别; 高斯混合–隐马尔科夫模型; 瓶颈特征; 深度神经网络

中图分类号 TP391

Study on Continuous Speech Recognition Based on Bottleneck Features for Lhasa-Tibetan Dialect

ZHOU Nan, ZHAO Yue†, LI Yaoqiang, XU Xiaona, CAIWANG Lamu, WU Licheng

School of Information Engineering, Minzu University of China, Beijing 100081;
† Corresponding author, E-mail: zhaoyueso@muc.edu.cn

Abstract The bottleneck features extracted from deep neural network not only have long term context-dependence and compact representation of speech signal, but also can replace the traditional MFCC features for GMM-HMM acoustic modeling. The authors apply bottleneck features and their concatenated features with MFCC into Lhasa-Tibetan continuous speech recognition. The experiments in Lhasa-Tibetan continuous speech recognition show that the concatenated features of bottleneck features and MFCC achieve better performance than the posterior features of deep neural network and mono-bottleneck features.

Key words Lhasa-Tibetan; continuous speech recognition; GMM-HMM; bottleneck features; deep neural network (DNN)

深度神经网络(deep neural network, DNN)是一种拥有很多隐层的机器学习模型, 该模型提取的深度特征具有分布平稳、表征能力强、易于建模等优点。目前深度学习理论已经成功地应用于大词汇量连续语音识别(large vocabulary continue speech recognition, LVCSR)中[1–2]。尽管深度学习在英文、中文等主要语种语音识别任务上的优势已得到实验验证, 但对其在藏语大词汇量连续语音识别任务中的应用只有少量研究。王辉等[3]和张宇聪[4]探讨了基于深度学习的藏语音素和藏语孤立词识别。袁胜龙等[5]研究了基于深度神经网络的藏语拉萨话连续语音识别, 将声韵母作为识别基元, 利用 DNN 输出层特性进行三音素 HMM 声学建模, 识别模型单音节识别率达到 43%左右。

在 DNN-HMM 声学模型应用之前, 传统的语音识别声学模型采用 MFCC (Mel-fequency cepstral coefficents, 梅尔频率倒谱系数)特征对 GMM-HMM 进行建模, GMM-HMM 模型具有完善的理论知识体系, 训练效率高, 但由于每帧的 MFCC 特征通常只包含毫秒级的语音信号, 信息量不足, 容易受噪声污染, 抗噪能力很弱[6]。

为了利用 GMM-HMM 的性能优势, 文献[7–9]

教育部人文社会科学规划基金(15YJAZH120)和中央民族大学双一流学科建设项目资助
收稿日期: 2017–05–31; 修回日期: 2017–09–05; 网络出版日期: 2017–11–06

研究了一种具有狭窄中间层的 BottleNeck (BN)深度神经网络, 提取网络中间的瓶颈特征来替代传统的 MFCC 语音特征, 用于训练传统的 GMM-HMM 模型。该特征不仅具有语音长时相关性, 而且具有高度抽象表征信号的能力。基于该思想, 本文采用瓶颈特征及其与 MFCC 复合特征的藏语连续语音识别技术, 验证瓶颈特征在藏语大词汇量连续语音识别中的有效性。

1 瓶颈特征提取

在语音信号处理过程中, 传统的声学特征参数容易受到环境噪声、说话人差异性以及信道的影响, 即使相同的语音内容, 在不同人发音时, 也会有很大差异。相对于底层声学参数, 高层信息具有更强的鲁棒性。具有狭窄中间层的 BottleNeck 神经网络是一种特殊结构的 DNN, 其对输入特征进行多次非线性变换, 得到区分性更强的声学特征, 模型中间层的 BN 特征可以替代传统的语音特征, 进行 GMM-HMM 声学模型建模。

提取 BN 特征需要先训练一个 DNN 神经网络, 步骤如下: 1) 运用对比散度算法, 训练受限玻尔兹曼机(restricted Boltzman machine, RBM), 得到参数。2) 将第一层输出作为第二层的输入, 训练第二层 RBM, 得到第二层的参数, 重复第二步, 直到达到所需深度。3) 采用有监督的训练算法, 利用误差反向传播算法来优化参数[10]。BN 层的结点比其他隐层少, 在 DNN 训练完成后, 将 BottleNeck 层的网络参数取出作为后续建模的瓶颈特征[11]。具体过程如图 1 所示。

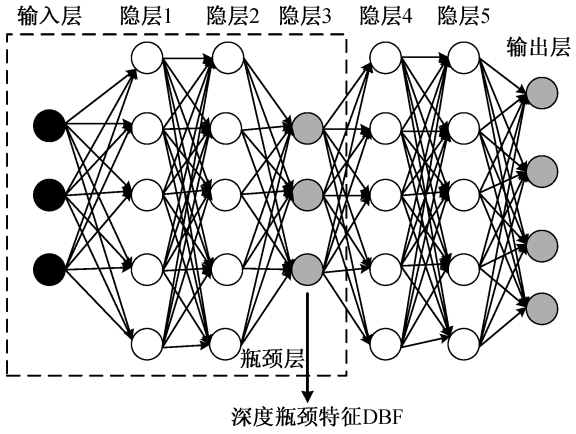

图 1 瓶颈结构深度神经组成结构
Fig. 1 BN-DNN composition structure diagram

1.1 RBM 实现深度学习

首先定义受限玻尔兹曼机 RBM。对于变量 $\boldsymbol{v}$ 和 $\boldsymbol{h}$, $\boldsymbol{v}=\left[v^1,v^2,...,v^n\right]^{\mathrm{T}}$, $\boldsymbol{h}=\left[h^1,h^2,...,h^m\right]^{\mathrm{T}}$, $v^i,h^j\in\{0,1\}$, $i=1,2,...,n$, $j=1,2,...,m$, 如果其分布函数为

$$p(\boldsymbol{v},\boldsymbol{h})=\frac{1}{Z}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}, \tag{1}$$

$$Z=\sum_{v,h}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}, \tag{2}$$

RBM 的能量函数 $E(\boldsymbol{v}, \boldsymbol{h})$表示为

$$E(\boldsymbol{v},\boldsymbol{h},\theta)=-\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}v_ih_j-\sum_{i=1}^{n}a_iv_i-\sum_{j=1}^{m}b_jh_j, \tag{3}$$

则符合式(3)分布的一组随机变量称为受限玻尔兹曼机。从式(3)可以看出, 可见结点的一组取值与隐藏结点的一组取值产生的概率 $p(\boldsymbol{v}, \boldsymbol{h})$与能量函数有相关性。由条件概率可得

$$p(\boldsymbol{v})=\frac{\sum_{h}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}, \tag{4}$$

$$p(\boldsymbol{h})=\frac{\sum_{v}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}, \tag{5}$$

$$p(\boldsymbol{v}\mid\boldsymbol{h})=\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum_{v}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}, \tag{6}$$

$$p(\boldsymbol{h}\mid\boldsymbol{v})=\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum_{h}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}。 \tag{7}$$

由此引入一个自由能量概念, 可以得到最大似然值。自由能量函数为

$$\mathrm{FreeEnergy}(\boldsymbol{v})=-\ln\sum_{h}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}, \tag{8}$$

则

$$p(\boldsymbol{v})=\frac{\mathrm{e}^{-\mathrm{FreeEngery}(\boldsymbol{v})}}{Z}。 \tag{9}$$

对式(9)两边取对数, 可得到

$$\ln p(\boldsymbol{v})=-\mathrm{FreeEngery}(\boldsymbol{v})-\ln Z。 \tag{10}$$

由前面的运算可以得到

$$\sum_{\boldsymbol{v}}\ln p(\boldsymbol{v})=\ln\left[\prod_{v}p(\boldsymbol{v})\right], \tag{11}$$

其中 $\ln\left[\prod_{v}p(\boldsymbol{v})\right]$是似然函数, 最大似然估计是使似

然函数最大。对于 RBM 系统，若要得到最大似然值，就是求其对应的参数 $\theta \in \{w_{ij}, a_i, b_j;\ i=1, 2, ..., n,\ j=1, 2, ..., m\}$。将 RBM 系统似然函数 $\prod_{\boldsymbol{v}} p(\boldsymbol{v})$ 写成下式:

$$L(\theta \mid \boldsymbol{v}) = \prod_{\boldsymbol{v}} p(\boldsymbol{v}), \tag{12}$$

对上式进行取对数并求偏导数，得到

$$\Delta w_{ij} = \varepsilon\left(\left\langle v_i h_j \right\rangle_{\text{data}} - \left\langle v_i h_j \right\rangle_{\text{model}}\right)。 \tag{13}$$

$\left\langle v_i h_j \right\rangle_{\text{data}}$ 表示输入样本的自由能量期望值，$\left\langle v_i h_j \right\rangle_{\text{model}}$ 表示模型产生的样本数据的自由能量期望值。

为了让系统似然函数达到最大，可通过上述训练找出一个能让系统自由能量总和最小的参数。这样，通过 RBM 训练，能更充分地训练多层神经网络模型，从而得到一个最优模型[12]。

1.2　深度瓶颈特征提取

完成 DNN 的训练后，将瓶颈层以上的隐含层和输出层网络参数移除，得到深度瓶颈特征的提取函数 $g(*)$，提取形式为

$$\boldsymbol{y} = g(\boldsymbol{x}), \tag{14}$$

其中，$\boldsymbol{y} \in R^{D^l} = \left[y_1, ..., y_D\right]^{\mathrm{T}}$，$D^l$ 表示 BottleNeck 层的结点数，输入特征矢量 $\boldsymbol{x} = \left[x_1, ..., x_D\right]^{\mathrm{T}}$。瓶颈层的每个特征结点输出计算公式如下:

$$y_{d^l} = g(\boldsymbol{x} \mid \theta) = \sum_j w^l_{d^l, j} f\left(\sum_i w^{l-1}_{j,i} f\left(...+ f\left(\sum_d w^l_{d,i} + b^1_i\right)...\right) + b^{l-1}_j\right) + b^l_d, \tag{15}$$

$d^l = 1, ..., D^l$ 中的 l 表示 BottleNeck 的层数，$w^l_{i,j}$ 表示第 l 层的结点 i 与第 $l-1$ 层的结点 j 之间的网络权重参数，b^l_i 表示第 l 层的结点 i 的偏置，$f(*)$表示激活函数。

1.3　瓶颈复合特征

声学复合特征指将非短时差异特征与传统短时特征拼接后形成的新特征参数。吕丹桔等[13]将短时声学特征 MFCC 与多层感知器(multi-layer perception, MLP)产生的两类差异性特征拼接成新的特征参数，利用新特征流对 GMM-HMM 进行声学建模，在中文语音识别率上比单一特征有较大提升。然而，多层感知器属于浅层神经网络，不能提取深层的语音特征，本文将 DNN 提取具有长时性的 39 维瓶颈特征与传统的 39 维 MFCC 特征复合成 78 维的高维特征参数，通过线性区分分析(LDA)[14]进行降维，降维后的 39 维特征参数用于 GMM-HMM 声学建模。特征流复合过程如图 2 所示。

2　藏语拉萨话音素集合

藏语属于汉藏语系藏缅语族支，主要分布在中国西藏自治区、青海、甘肃、四川、云南等省份。全国大约有 600 多万人使用藏语。中国的邻国印度、巴基斯坦、尼泊尔等国家也有藏语分布。藏语有 3 种方言: 卫藏、康巴和安多，三者书面语相同，发音不同[15]。

藏语属于拼音文字，音节是藏语的基本单位，由一个或几个音素按一定的规律组合而成，音节之间用音节符(“·”)分隔[16]。藏文具有与英文一样的书写规则，整体上自左向右横向书写，但藏文的一个

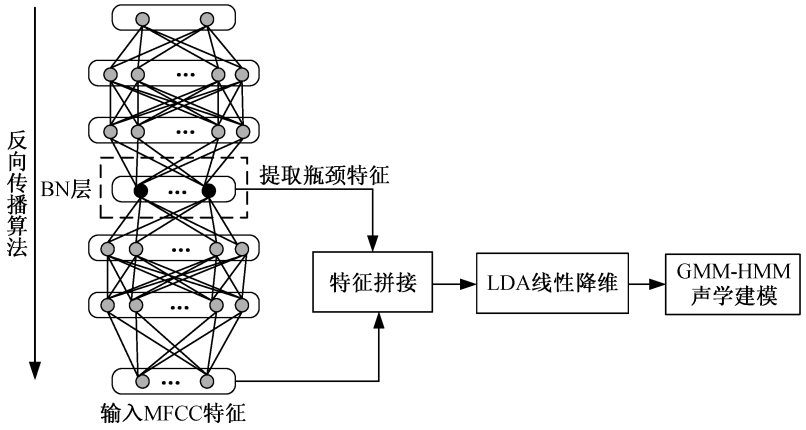

图 2　瓶颈复合特征提取流程

Fig. 2　Extraction of concatenated bottleneck feature

音节内的书写有竖向叠加，称为“叠加书写”，叠加书写结构中以一个辅音字母为中心位置，将字母分为“前加字”、“上加字”、“下加字”、“后加字”和“再后加字”，中心位置的辅音字母称为“基字”(图 3)。

藏语共有 30 个辅音字母和 4 个元音符号，每个辅音字母都可以做为音节的基字，4 个元音符号中，“ི”、“ུ”和“ེ”以上加的形式与辅音组合，“ོ”以下加的形式与辅音组合，用于改变辅音字母的母音发音。藏语的发音顺序为

音节=前加字+不带元音的基字丁+下加字+元音+后加字+再后加字[17]。

藏语拉萨话定义为藏语的标准发音。在发音上，拉萨话的声母含有复辅音以及单辅音，一般常用单辅音声母。拉萨话的韵母由核心元音与韵尾两部分组成。由于藏语是按一字一音准确拼写的字母拼音文字，因此理论上藏语语音识别系统采用音素基元比较科学[15]。本文将拉萨话的声韵母拆分成

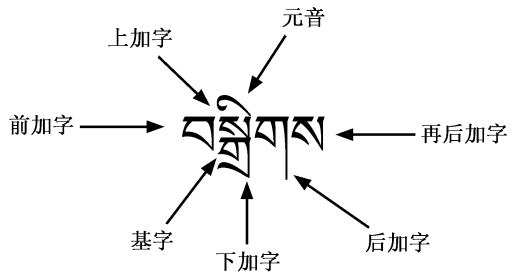

图 3 藏文的音节结构

Fig. 3 Syllable structure of Tibetan language

颗粒度更小的音素，整理后的音素集包含 59 个音素，并制订了相应的拉丁转写方案，如表 1 所示。

3 实验与结果分析

3.1 语料库与评价指标

本文的语料库收集自藏语拉萨话发音标准的 8 个女生和 8 个男生的日常口语语音。选用的句子每句长度为 2~60 秒，形成约 10 小时的藏语连续语音语料库。录制过程中，采用 16 KHz 采样频率、16 bit 的量化精度。语料中的 8.5 小时用于模型训练，1.5 小时做测试。

实验评价指标为藏语连续语音识别中的音节错误率(WER)。设 N 为测试文本中词的数量，I 为插入词个数，D为删除词个数，S为替代词个数。WER 的定义为

$$\mathrm{WER}=\frac{I+D+S}{N}。\tag{16}$$

3.2 语言模型

由于语言模型在整个语音识别中非常重要，因此有无语言模型对整个语音识别系统的使用范围和识别效率影响很大。目前，大部分连续语音识别系统均采用基于 N-gram 的统计语言模型，用于在解码器中计算语言模型的概率[18]。本文使用训练集中的文本来训练词的三元语言模型，并采用回退平滑技术来解决数据稀疏的问题。

搭建藏语语音识别系统时，使用开源工具箱 kaldi[19]进行相关实验，包括数据准备、声学特征提

表 1 藏语拉萨话音素及拉丁转写表

Table 1 Lhasa-Tibetan dialect phonemes and Latin Transliteration table

IPA	拉丁转写	IPA	拉丁转写	IPA	拉丁转写	IPA	拉丁转写	IPA	拉丁转写
c	gy	nc	jy	c^h	ky	h	h	a:	v
ʂ	hr	j	y	k	k	ŋk	gh	k^h	g
l	l	ɬ	lh	m	m	n	n	ŋ	ng
ɲ	ny	p	p	mp	bh	p^h	b	r	r
t	t	nt	dh	ts^h	tsh	tʂ	ts	nts^h	th
$tʂ^h$	khr	t^h	d	tɕ	c	ntɕ	jh	$tɕ^h$	qj
br	br	ntʂ	z	w	w	ɕ	sh	f	ff
a	a	ʔ	ab	e	e	eʔ	eb	e:	ew
$\tilde{e}$	eu	ɛ	el	ɛʔ	elb	ɛ:	elw	Øʔ	fb
Ø	f	Ø:	fw	i	i	iʔ	ib	$\tilde{i}$	il
i:	jw	o	o	o:	ow	u	u	u:	uw
y:	yw	yʔ	yb	$\tilde{y}$	yu	s	s		

说明: IPA 是国际音标 International Phonetic Alphabet 的缩写。

取、声学模型的训练与解码，使用 DNN 工具包进行 DNN 相关的搭建与训练。

3.3　GMM-HMM 模型

本文使用三音素建模。对普通的三音素单元，使用自左向右的无状态间跨越的三状态 HMM；对静音模型，则采用状态间可跨越的五状态 HMM，每个 HMM 拓扑结构前后都有一个开始状态和一个结束状态。结合对藏语拉萨话语音发音特征的研究成果[20]，我们将拉萨话的 59 个音素划分为 20 个类别集，并分别建立决策树问题集。

GMM-HMM 模型用最大似然估计准则来训练，输入是 39 维特征，帧长为 25 ms，帧移为 10 ms。HMM 中每个状态设置 100 个独立的高斯分量。GMM-HMM 模型结构如图 4 所示。

3.4　BN-GMM-HMM 模型

传统的瓶颈特征由多层感知器获取，采用的是一种非线性的特征变换和降维的技术[21]。本文基于 DNN 模型生成瓶颈声学特征，DNN 的训练特征使用 39 维的 MFCC 特征(12 维滤波器输出值加上 1 维对数能量，以及其一阶差分和二阶差分)，并对当前帧进行前后各 5 帧拼接，共 11 帧的上下文扩展。BN-DNN 网络总共 7 层，包含 1 个输入层、5 个隐含层和 1 个输出层，BN 层定义在中间一个隐层。输入层包含 429 个结点，输出层包含 1153 个结点，每个结点对应不同的三音素。BN 层结点数与 MFCC 特征维数一样，有 39 个结点，其他隐含层分别包含 1024 个结点。因此，BN-DNN 的结构为“429-1024-1024-39-1024-1024-1153”。对每个隐含层进行 100 次 RBM 预训练，然后利用 BP 算法进行全局参数的微调，将预估概率分布之间的交叉熵作为目标函数。在训练过程中，模型初始学习速率设置为 0.08，在前 4 个迭代训练过程中保持不变，后续训练时将学习速率减半，当交叉验证值趋于收敛时停止训练。此外，将冲量值设为 0.5，Mini-batch 的大小设为 256。训练完成后，利用 BN 特征训练得到的 GMM-HMM 模型进行识别解码。

3.5　BN+MFCC-GMM-HMM 模型

BN+MFCC-GMM-HMM 模型用提取的 BN 特征串接原始 MFCC 特征，通过 LDA 降维，训练得到 GMM-HMM 模型，并进行识别解码。训练涉及的参数配置与 BN-DNN 训练一致。

3.6　实验结果及分析

在 Kaldi 上分别搭建基于 MFCC、BN 和 BN＋MFCC 特征的 GMM-HMM 声学模型语音识别系统以及 DNN-HMM 声学模型系统，对藏语进行语音识别性能测试。将音节错误率作为系统性能的衡量指标，结果如表 2 所示。

表 2 显示瓶颈复合特征、DNN 后验特征以及单一瓶颈特征。与基线模型 MFCC-GMM-HMM 相

表 2　模型识别结果

Table 2　Recognition results of models

模型类别	WER/%
MFCC-GMM-HMM	18.61
DNN-HMM	15.77
BN-GMM-HMM	15.98
BN+MFCC-GMM-HMM	14.49

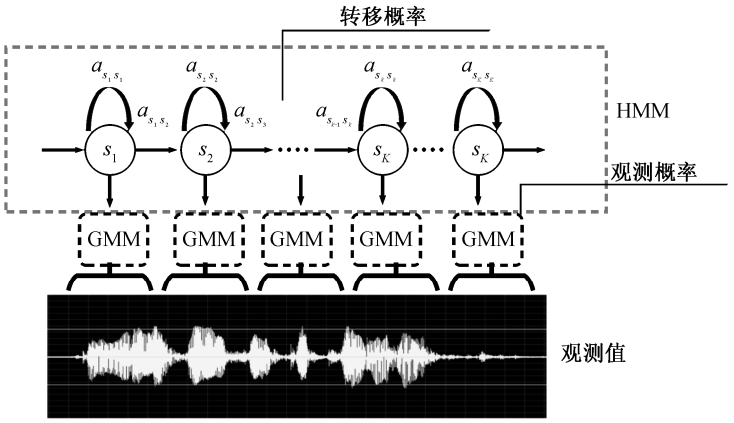

图 4　高斯混合–隐马尔科夫声学模型

Fig. 4　GMM-HMM Model

比, 藏语连续语音识别音节错误率分别下降4.12%, 2.84%和2.63%。瓶颈复合特征具有最佳的识别表现, 单一的瓶颈特征和深度神经网络后验特征的识别表现相当。瓶颈特征不仅能够借助成熟的 GMM-HMM 技术训练声学建模, 而且融合传统 MFCC 特征后, 系统识别率得到更加明显的改善。

4 结语

本文研究了瓶颈特征及其复合特征在藏语拉萨话连续语音识别中的应用问题, 并在 Kaldi 系统平台上进行实验。结果显示, 与传统 MFCC 特征相比, BN 特征在藏语语音识别率上更有优势, 而其复合特征与 DNN 后验特征和单 BN 特征相比, 能够进一步降低藏语连续语音识别音节错误率。瓶颈复合特征不仅结合了 DNN 特征语音长时相关性和高度抽象表征信号的能力, 还结合了 MFCC 特征具有人耳听觉的特性, 利用了成熟的 GMM-HMM 声学建模技术, 在藏语拉萨话连续语音识别方面表现得更好。将来, 我们会加大训练语料的时长, 采用不同神经网络模型进行训练比对, 以期获得更高的系统识别率。

参考文献

[1] Abdel-Rahman M, Yu D, Deng L. Investigation of full-sequence training of deep belief networks for speech recognition // Eleventh Annual Conference of the International Speech Communication Association. Makuhari, 2010: 2846–2849

[2] Dahl G E, Yu D, Deng L, et al. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, 2011: 4688–4691

[3] 王辉, 赵悦, 刘晓凤, 等. 基于深度特征学习的藏语语音识别. 东北师大学报: 自然科学版, 2015, 47(4): 70–73

[4] 张宇聪. 基于深度学习的藏语拉萨方言语音识别的研究[D]. 兰州: 西北师范大学物理与电子工程学院, 2016

[5] 袁胜龙, 郭武, 戴礼荣. 基于深层神经网络的藏语识别. 模式识别与人工智能, 2015, 28(3): 210–213

[6] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 2006, 13 (5): 308–311

[7] Yu D, Seltzer M. Improved bottleneck features using pretrained deep neural networks // Twelfth Annual Conference of the International Speech Communication Association. 2011: 237–240

[8] 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. 数据采集理, 2015, 30(2): 365–371

[9] 刘学, 王松年, 郭武. 采用深层神经网络中间层特征的关键词识别. 小型微型计算机统, 2015, 36(7): 1541–1544

[10] Geoffrey H, Deng L, Yu D, et al. Deep neural Networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82–97

[11] Song Y, Jiang B, Bao Y B, et al. I-vector representation based on bottleneck features for language identification. Electronics Letters, 2013, 49(24): 1569–1570

[12] Geoffrey H. A practical guide to training restricted boltzmann machines // Neural Networks: Tricks of the Trade. Berlin: Springer, 2012: 599–619

[13] 吕丹桔, Hoffmeister B. 汉语语音声学特征复合的研究. 云南大学学报: 自然科学版, 2010, 32(增刊 1): 368–371

[14] Gopinath R A. Maximum likelihood modeling with Gaussian distributions for classification // IEEE International Conference on Acoustics, Speech and Signal Processing. Seattle, 1998: 661–664

[15] 李冠宇, 孟猛. 藏语拉萨话大词表连续语音识别声学模型研究. 计算机工程, 2012, 38(5): 189–191

[16] 金鹏. 中国少数民族语言简志丛书藏语简志. 北京: 民族出版社, 1983

[17] 拉龙东智. 藏语语音识别技术研究[D]. 拉萨: 西藏大学藏文信息技术研究中心, 2015

[18] Ablimit M, Neubig G, Mimura M, et al. Uyghur morpheme-based language models and ASR // IEEE 10th International Conference on Signal Processing (ICSP). Beijing, 2010: 581–584

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

[20] Zhao Y, Zhao R, Wang Xiaoyang, et al. Multilingual articulatory features augmentation learning // 23rd International Conference of Pattern Recognition. Cancun, 2016: 2902–2906

[21] Grézl F, Karafiát M, Kontár S, et al. Probabilistic and bottleneck features for LVCSR of meetings // IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu, 2007: IV-757–IV-760