计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(3):29−38 [doi: 10.15888/j.cnki.csa.007315]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

藏语语音识别研究进展和展望①

王福钊[1], 周　雁[2]

[1](西藏大学 信息科学技术学院, 拉萨 850000)
[2](北京理工大学珠海学院 计算机学院, 珠海 519088)
通讯作者: 周　雁, E-mail: 37942264@qq.com

摘　要: 随着英汉语音识别技术的不断发展, 对少数民族语言语音识别技术的研究也紧跟其后并取得了一定的成果. 藏族人民是中华民族大家庭中不可或缺的一员, 藏语语音识别技术研究是语音识别技术研究中不可缺少的重要部分. 文章首先对国内藏语言语音识别的研究历程及研究改进之处进行了梳理, 其次从藏语言本身的文字特点以及发音特点和要素出发详细介绍了藏语语音识别研究中使用到的基于模板匹配、统计概率模型以及人工神经网络 3 种方法, 并对 3 种方法各自的特点和适用范围进行了总结归纳, 最后从藏语言语音识别研究进展和各识别方法的自身特点出发探讨了语音识别研究中存在的难点问题, 并展望了其未来发展的方向.

关键词: 语音识别; 藏语; 模板匹配; 统计概率; 神经网络; 研究展望

引用格式: 王福钊,周雁.藏语语音识别研究进展和展望.计算机系统应用,2020,29(3):29–38. http://www.c-s-a.org.cn/1003-3254/7315.html

Progress and Prospects of Tibetan Speech Recognition Research

WANG Fu-Zhao[1], ZHOU Yan[2]

[1](School of Information Science and Technology, Tibet University, Lhasa 850000, China)
[2](School of Computer Technology, Beijing Institute of Technology, Zhuhai, Zhuhai 519088, China)

Abstract: With the continuous development of English and Chinese speech recognition technology, the research on minority language speech recognition technology has followed closely and achieved certain results. The Tibetan people are an indispensable member of the Chinese nation's family. The study of Tibetan speech recognition technology is an indispensable part of the research of speech recognition technology. Firstly, the paper presents the research process and research improvement of Tibetan speech recognition in China. Secondly, it introduces the template-based matching and statistical probability model and artificial neural network used in Tibetan speech recognition research from the characteristics of Tibetan language itself and its pronunciation features and elements, then summarizes the characteristics and application scope of the three methods. Finally, it discusses the research progress of Tibetan speech recognition and the characteristics of each recognition method, discusses the difficult problem and the direction of its future development.

Key words: speech recognition; Tibetan; template matching; statistical probability; neural network; research prospect

从狭义上讲, 自动语音识别 (Auto-Speech Recognition, ASR) 是指将人类语音信号自动转换成相应的文本内容的机器程序执行过程. 但就其广义上讲, 语音识别是指能够理解人类的语音信号的机器程序执行过程, 包括声纹理解和内容理解. 语音识别是一门跨学科技术, 结合了物理声学、语言学、信号处理学、生理

① 基金项目: 西藏自治区自然科学基金 (XZ2019ZRG-09)
Foundation item: Natural Science Foundation of Tibet Autonomous Region (XZ2019ZRG-09)
收稿时间: 2019-07-27; 修改时间: 2019-09-02; 采用时间: 2019-09-18; csa 在线出版时间: 2020-02-28

学、计算机学和统计概率学. 语音识别研究可追溯到 20 世纪 50 年代, 贝尔实验室成功研究实现了 Audry 语音识别系统[1]. 自 20 世纪 80 年代以来, 国内汉语语音识别研究取得了空前的发展. 到目前, 以百度、科大讯飞为首的公司已将深度学习神经网络成功运用在汉语普通话和各方言的语音识别上, 识别效果好, 识别率高, 识别技术成熟.

藏族是中华民族大家庭中重要的一员, 其人口近 630 万人 (来自第 6 次人口普查数据), 主要分布在我国康巴地区 (西藏、四川、云南)、安多地区 (西藏、甘肃、青海、四川) 和卫藏地区 (西藏). 藏语是藏族人民交流和沟通的主要载体. 藏语起源可追溯至上古象雄语言学时期, 其正式成文可追溯至七世纪吐蕃王朝松赞干布时期, 后经中世纪和近代语言学时期的发展形成了如今这具有统一的文字、语法、字法、词法以及书写法的独特魅力语言[2].

随着信息化和智能化时代的到来, 藏语言语音识别具有不可替代的重要作用. 首先, 藏语言语音识别是藏区智能化发展过程中不可或缺的关键技术. 第二, 藏语言语音识别对加强藏区内各地区 (卫藏、康巴和安多)、藏区与其他地区的交流和沟通, 进而有效加强民族融合、增进民族和谐方面具有突出的重要地位和意义.

1 藏语语音识别研究现状

藏语语音识别是在英语、汉语的语音识别研究基础上, 从物理声学、语言学角度出发结合藏语自身特点采用计算机技术实现自动语音转换文本. 国内藏语语音识别的研究始于本世纪初, 滞后于汉语语音识别. 经过十多年的快速发展, 藏语语音识别研究取得了一定的成果.

2006 年, 李洪波和于洪志研究了基于藏文音节和文字特性的藏语语音识别基元, 选择以音素为识别基元进行语音端点检测, 从而提高了噪音背景下语音识别效率[3].

2007 年, 西北民族大学于洪志、李永宏、索南楞次等研究创建了安多藏语单音节文本库、语音库和声学参数数据库, 并针对单音节的语音声学特征进行了较为系统的研究[4]; 同年, 刘静萍和德熙嘉措通过提取 LPCC 参数并使用 DTW 实现了安多藏语小词汇孤立词语音识别系统[5]; 同年, 武光利、戴玉刚等通过短时平均幅度和短时平均过零率相结合的方法来改进了藏语语音端点检测技术[6]; 同年, 李洪波和于洪志通过提取 MFCC 参数进行语音识别而提高了识别效率[7].

2009 年, 李勇、于洪志、达哇彭措研究提取了藏语语音韵律特征用于语音识别进而提高了语音识别效率[8]; 同年, 姚徐、李永宏、单广荣等提取了 MFCC 特征参数, 并构建了语音模板库, 采用 DTW 技术实现了藏语语音识别系统[9].

2010 年, 西藏大学德庆卓玛以声韵母作为识别基元, 分别提取了 LPCC 和 MFCC 特征参数实现拉萨话藏语特定人小词汇量语音识别系统, 并比较了两者的识别效果[10]; 同年, 韩清华改进了语音端点检测, 并对提取的 MFCC 参数进行矢量量化, 采用 HMM 进行声学建模实现了藏语安多方言非特定人孤立词语音识别系统[11].

2011 年, 西南交通大学刘巧凤引入快速沃尔什变换对 MFCC 提取进行改进, 进而提高了藏语语音识别的特征有效性和性能优越性[12].

2012 年, 李冠宇分别以音素和声韵母作为识别基元, 通过 HTK 工具包构建了一个上下文相关的拉萨话藏语大词量连续语音识别声学模型[13].

2015 年, 赵尔平对传统的特征提取方法进行了改进, 在 MFCC 特征向量的基础上结合拉萨话语音特点融入共振峰参数, 提高了拉萨话藏语孤立词的语音识别率[14]; 同年, 中央民族大学许彦敏引入了种子模型建立了藏语单音素和三音素声学模型, 并实现了基于 sparse auto-encode 的英藏跨语言语音识别系统[15]; 同年, 王辉、赵悦、刘晓凤等将提取到的 MFCC 特征提取使用稀疏自动编码器提取了语音深度特征, 进而实现了基于深度特征的藏语语音识别系统[16].

2016 年, 中央民族大学刘晓凤比较了基于 MFCC 特征参数、SA+MFCC 特征参数和 DBN+MFCC 特征参数的藏语连续语音识别率, 提出了通过 DBN 模型处理 MFCC 提取的深度特征对藏语连续语音识别有更高的识别效率[17]; 同年, 西北师范大学张宇聪利用长短时记忆网络模型提取深度特征, 实现了基于深度学习的拉萨话藏语语音识别系统[18].

2017 年, 中央民族大学周楠将深度神经网络提取的瓶颈特征与传统 MFCC 组合形成了复合特征. 并对提取的复合特征和单瓶颈特征进行了在藏语语音识别

率上的比较[19].

2018 年, 中央民族大学赵悦、李要嫱、徐晓娜等采用基于主动学习的语音语料选取方法实现了利用少量语音样本构建了能够代表大量语音样本的高精度藏语拉萨话识别模型, 成功降低了语音预处理过程中语音语料人工标注的工作复杂度[20]. 同年, 梁宁娜、邓彦松其中在传统双门限检测法的基础上进行了端点放松处理对端点检测技术进行了改进, 再采用 DTW 实现藏语孤立词语音识别, 实验成功的提高了在噪声下的语音识别效率[21]. 同年, 中科大的黄晓辉、李京利用 RNN 和连续时序分类算法实现了端到端的藏语语音声学建模[22]. 同年, 陕西师范大学李涛、曹辉、郭乐乐通过堆叠稀疏自编码器组成深度自编码器提取了深度特征并实现了基于深度特征的藏语语音识别[23].

2 藏语概述

2.1 藏文结构

藏文类似于汉文属于拼音型文字. 从狭义上讲, 藏文是指藏语的符号; 但就广义上讲, 藏文除了符号外还包括藏文文法等. 藏文音节是现代藏文文本的基本组成单位, 藏文音节由 30 个辅音字母和 5 个元音字母(其中ཨa 为省略不写) 组成[2]. 如表 1 及表 2 所示.

表 1 藏文辅音字母及其拉丁转写

组名	组员 1	组员 2	组员 3	组员 4	组名	组员 1	组员 2	组员 3	组员 4
ཀ་མདེད།	ཀ k	ཁ kh	ག g	ང ng	ཅ་མདེད།	ཅ c	ཆ ch	ཇ j	ཉ ny
ཏ་མདེད།	ཏ t	ཐ th	ད d	ན n	པ་མདེད།	པ p	ཕ ph	བ b	མ m
ཙ་མདེད།	ཙ ts	ཚ tsh	ཛ dz	ཝ w	ཞ་མདེད།	ཞ zh	ཟ z	འ v	ཡ y
ར་མདེད།	ར r	ལ l	ཤ sh	ས s	ཧ་མདེད།	ཧ h	ཨ a		

表 2 藏文元音字母及其拉丁转写

ི	ུ	ེ	ོ
i	u	e	o
/i/	/u/	/e/	/o/

藏文音节 (也称藏字) 间用隔音符“”隔开, 句子间用单垂符“”或双垂符“||”隔开. 藏文音节是横纵双向叠加的平面字形, 并由基字、前加字、上加字、下加字、元音符号、后加字和再后加字中的 1~7 部分组成. 字形结构如图 1 所示.

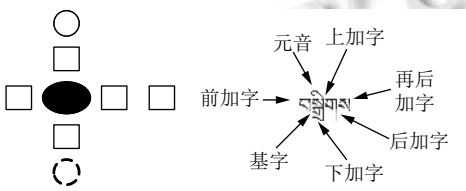

图 1 藏文音节结构

图 1 中, 基字: 30 个辅音字母皆可作为基字.

前加字: ག ད བ མ འ共 5 个.

上加字: ར ལ ས共 3 个.

下加字: ཡ ར ལ ཝ共 4 个.

元音符号: ི ུ ེ ོ共 4 个.

后加字: ག ང ད ན བ མ འ ར ལ ས共 10 个.

再后加字: ད ས共 2 个.

藏字在结构上有严格的规则限制. 基字是组成藏字不可缺少的构件, 其余各构件通过构字规则进行约束出现.

2.2 藏语发音

藏语发音过程是通过声带按照音节拼读规则振动产生声音的过程. 在藏语表达过程中, 其声音音素携带了语音信号的大量信息. 藏字发音时基字发主音, 前加字、上加字、下加字、后加字、再后加字发辅助音. 藏字的拼读发音顺序为前加字→不带元音的基字丁(不带元音符号的纵向叠加部分)→元音→后加字→再后加字[16].

值得注意的是藏语三大方言虽文字相同, 但发音存在较大差异. 具体差异如下: (1) 安多方言没有声调、复元音、长元音, 而卫藏和康巴方言有该特征, 且安多方言复辅音比较丰富; (2) 卫藏方言没有送气清擦音、清化鼻音、浊塞擦音和浊塞音, 而安多和康巴方言则具有, 且卫藏方言具有单辅音和复辅音; (3) 康巴方言和安多方言很少有舌面擦音 c 和 ch, 但卫藏方言中有[24]; (4) 藏语拉萨话在发音上声母发音不含浊音和塞音, 这与汉语的发音不同. 复辅音声母出现较少, 有真性复合和鼻化元音, 声调起伏波动不大, 发音也较为

平稳[25].

藏字发音携带了特定的音素信息, 故可以通过对音素进行特征提取来表达语音信号的内容. 在藏语拉萨话中有 59 个音素[13]. 如表 3 所示.

表 3 藏语拉萨话音素及其拉丁转写

音素	拉丁转写	音素	拉丁转写	音素	拉丁转写	音素	拉丁转写	音素	拉丁转写
c	c	kh	kh	ph	ph	ŋ̊	nn		tx
ch	ch	l	l	r	r	e:	ew	o	o
h	h	m	m	s	s	eˀ	eb	o:	ow
j	j	n	n	t	t	ɪ	i	u	u
k	k	p	p	th	th	i:	iw	u:	uw
tʂ	q	ɕ	x	tɕh	txh	iˀ	ib	y:	yw
txh	qh	ʂ	ss	ʔ	ab	ɛ	el	ɛ:	elw
w	w	ʦ	ts	a:	aw	ɛ̃	eu	ỹ	yu
ŋ	ng	tsh	tsh	a	a	e	e	yˀ	yb
ø	f	ø:	fw	ĩ	il	ɛˀ	elb		

藏语同汉语一样可以将一个音节分离成声母和韵母来表示. 藏语拉萨话包含了 36 个声母和 45 个韵母[1]. 藏语拉萨话声、韵母及其拉丁转写如表 4、表 5 所示.

表 4 藏语拉萨话韵母、音标及拉丁转写

韵母	音标	拉丁转写	韵母	音标	拉丁转写	韵母	音标	拉丁转写
ཨ	a	ea	ཨེན	ẽ	en	ཨོ	o	eo
ཨག་ཨགས	ak	ak	ཨེར	er	er	ཨོབ་ཨོབས	ob	op
ཨམ་ཨམས	am	am	ཨེལ	e:	el	ཨོག་ཨོགས	ok	ok
ཨང་ཨངས	aŋ	ag	ཨོད་ཨོས	ɸ?	od	ཨོམ་ཨོམས	om	om
ཨབ་ཨབས	ap	ap	ཨོན	ɸ	on	ཨོང་ཨོངས	oŋ	og
ཨར	ar	ar	ཨོལ	ɸ:	ol	ཨོར	or	or
ཨེ	e	ee	ཨི	i	ei	ཨུ	u	eu
ཨེག་ཨེགས	e?	ek	ཨིད་ཨིས	i?	id	ཨུག་ཨུགས	uk	uk
ཨེད་ཨེས	ʔ	ed	ཨིན	ĩ	in	ཨུམ་ཨུམས	um	um
ཨན	ɛ	an	ཨིམ་ཨིམས	im	im	ཨུང་ཨུངས	uŋ	ug
ཨད་ཨས	ɛ?	ad	ཨིང་ཨིངས	iŋ	ig	ཨུབ་ཨུབས	ub	up
ཨེམ་ཨེམས	im	em	ཨིབ་ཨིབས	up	ip	ཨུར	ur	ur
ཨེང་ཨེངས	eŋ	eg	ཨིར	ir:	ir	ཨུད་ཨུས	ed	ud
ཨེབ་ཨེབས	ep	ep	ཨིལ	i:	il	ཨུན	un	un
ཨལ	ɛ:	al	ཨིག་ཨིགས	uk	ik	ཨུལ	ul	ul

3 藏语语音识别技术

藏语语音识别过程一般包括几个重要阶段: 语音数字化、预处理、特征参数提取、模型训练和模式匹配. 其原理如图 2 所示.

(1) 语音采集及处理. 常使用 CoolEdit 等工具采集一定频率、声道和分辨率的语音. 对采集的语音通过语音增强技术进行去噪.

(2) 语音信号预加重. 由于语音由声道产生后从嘴唇发出, 此过程中受口腔辐射影响会有高频损失, 为弥补这些高频信号损失, 常使用高通数字滤波器来增强高频语音信号[26]. 滤波器传递函数如式 (1) 所示.

$$H(z) = 1 - \alpha z^{-1}, 0.9 \le \alpha \le 1.0 \quad (1)$$

(3) 语音端点检测. 语音端点检测是用于检测语音信号段和非语音信号段. 一般使用基于短时能量和短时平均过零率的端点检测方法端点方法.

(4) 语音特征参数提取. 藏语的元音比辅音携带了

更多听觉感知信息. 通常提取语音信号的 LPCC (线性预测倒谱系数)、MFCC (Mel 频率倒谱系数)、PLPC (基于听觉模型的感知线性预测) 和深度特征等. 由于 Mel 频率更符合人耳听觉感知, 故一般提取 MFCC 特征用于语音识别. 声音频率和梅尔频率之间的转换关系如式 (2) 所示.

$$mel = 2595 \times \log 10(1 + freq/700) \quad (2)$$

随着语音技术的发展, 现在已经有更多的研究开始提取深度特征来代替原来的 MFCC 特征.

(5) 训练与识别. 模式创建过程即是模型训练的过程, 语音识别过程即是模式匹配的过程. 藏语语音识别技术研究至今已经有较大的进展, 从提取的不同特征和识别原理上, 可将藏语语音识别技术分为基于模板匹配的藏语语音识别技术、基于统计概率模型的藏语语音识别技术和基于人工神经网络的藏语语音识别技术 3 类.

表 5 藏语拉萨话声母、音标及拉丁转写

声母	音标	拉丁转写	声母	音标	拉丁转写
ཀྱ་དཀྱ་བཀྱ་རྐྱ་སྐྱ་བརྐྱ་བསྐྱ་དགྱ་བགྱ་རྒྱ་སྒྱ་བརྒྱ་བསྒྱ	c	jg	ཕ་འཕ་བ	ph	pa
མཉྱ་འཉྱ	nc	jy	དབ་ར་ར །	r	ra
ལྕ་མལྕ་འལྕ་ཀྱ	ch	qa	ཟ་གཟ་བཟ་ས་ས་་གས་བས་བས་	z	sa
ཧ་ཧ །	H	ha	ཏ་གདཏ་བདཏ་རྟ་ལྟ་སྟ་བསྟ་དབ་ར་བསབ་བསབ་བའ་ད་ཟའ་བཟའ	N	da
འ་ཨ	?	aa	མདྲ་འདྲ	nt	dh
ཤྲ	ş	hr	ཚཚ ། མཚ་འཚཚ	tsh	ca
དབ་ཡ་གཡ	j	ya	ཀྱ་དཀྱ་བཀྱ་སྐྱ་བསྐྱ་ཙ་རཙ ། གཙ་བཙ་རཙ་ཚ་ཙ་བཙ་རཙ	tş	td
ཀ་དཀ་བཀ་རྐ་ལྐ་སྐ་བརྐ་བསྐ་དག་བག་རྒ་ལྒ་སྒ་བརྒ་བསྒ	k	ga	མཇ་འཇ	ntsh	th
མཁ་འཁ	ŋK	gh	ལྕ་མལྕ་འལྕ་གྱ་གྱ ། ཚ་འཚ་	tşh	tf
ལ་ལ ། མལ་འལ་ག	Kh	ka	ཐ་མཐ་འཐ་ད	th	ta
ཀྱཚ་བཀྱཚ་གྱཚ་རྩཚ་བདྲཚ་ལ་ལ །	l	la	ཅ་གཅ་བཅ་ལྕ་རྩ་དབ་ས་ཤ་	tɕ	ja
སྨཚ་ལ་ཏྲ་བསྨཚ	lh	lh	ལྡ་བདྲ་མཇ་འཇ་འབ་	ntɕ	jh
མ་དམ་རྩ་སྨ་སྨ་	m	ma	ཚམཚ་འཚ་ཇ་ཤ་ ། འཚ་བ་	tɕh	qj
ན་གན་མཁ་རྩ་སྨ་བཏ་བདྲ་བསྨ	N	na	དག་་བག་་སྨ་་བསྨ་་ཀ་་ད་་ཡ་་དཡ་་སྨ་་བ་་སྨ་	br	zh
ང་དང་མང་རྩ་ལ་ཏ་སྨ་བདྲ་བསྨ	ŋ	ng	འདྲ་འབ་མཉ་འཉ་	nts	za
ཨ་ཨ ། གཨ་མཨ་རྩ་ཨ་བཨ་བསྨ་མ་རྩ་ཨ་	ɲ	ny	དབ་ཡ	w	wa
པ་དཔའ་ལ་ས་རྩ་ལ་ཏ་སྨ	p	ba	ཁ་ཁ ། གཁ་བཁ་ཤ་ཤ ། གཤ་བཤ	z	xa
འབ	mp	bh	ཀ་	f	fa

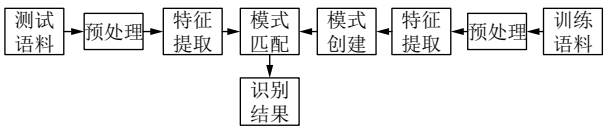

图 2 语音识别原理图

3.1 基于模板匹配的藏语语音识别技术

基于模板匹配的藏语语音识别技术是应用最早且、最成熟的藏语语音识别技术. 模板匹配的方法, 顾名思义就是通过创建参考模板库, 再由测试模板与模板库中的参考模板一一进行比较, 输出具有最大相似度的参考模板相对应的文本作为识别结果. 其原理图如图 3 所示.

基于模板匹配方法最常用的方法是动态时间规整 (Dynamic Time Warping, DTW), 其主要用来计算两个不等时长语音之间的相似度. 在模板训练方面, 有针对特定人群小词汇量的偶然模板训练法、有针对特定人群大词汇量的多模板平均训练法和针对非特定人群的聚类训练法[10]. 在系统实现中根据实际应用场景进行

模板训练方法选择. 在语音识别上, 通过计算欧几里德距离获得测试模板和参考模板之间的相似度, 并将相似度最大的作为识别结果输出. 欧氏距离计算公式如式 (3) 所示.

$$D[T(n),R(m)]=\sum_{n=1}^{p}(t_n-r_n)^2 \tag{3}$$

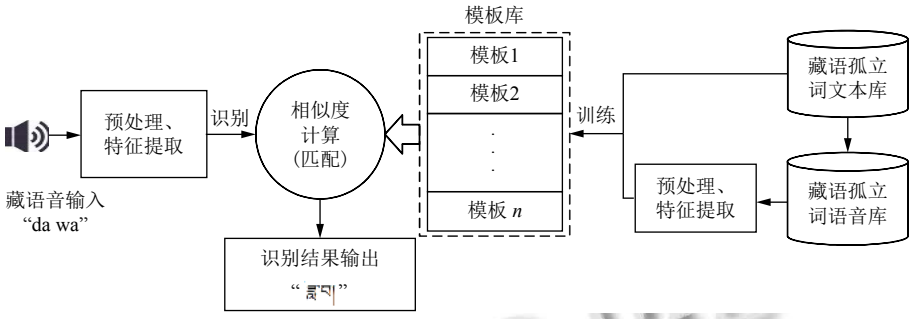

图 3 基于模板库匹配的藏语语音识别方法原理

3.2 基于统计概率模型的藏语语音识别技术

基于统计概率模型的藏语音识别方法是通过统计概率学知识构建训练语料语音音素序列的概率网络, 再根据该概率网络找到测试语料语音的可能音素序列, 从而实现语音识别.

利用 GMM-HMM 训练提取的 MFCC 参数建立声学模型, 通过 *N*-gram 方法建立语言模型. 对于藏语声学建模过程而言, 首先, HMM 状态序列是由藏语音节发音过程中选取的音素经过一系列的过程变化构成的. 其次, 观测向量 (即 MFCC 特征向量) 是由每一个音素以一定的概率密度函数生成. 最后, 使用高斯混合函数来拟合这种概率密度函数来表示具有随机特性的语音信号[24]. 基于统计概率模型的藏语语音识别技术原理图如图 4 所示.

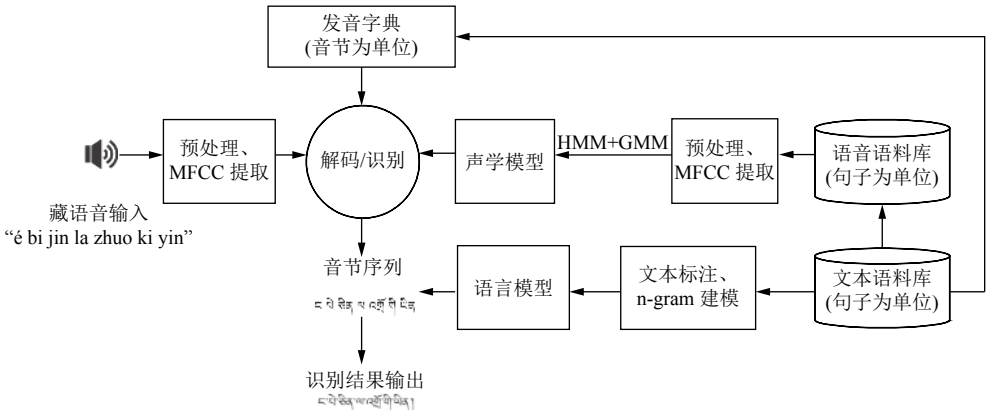

图 4 基于统计概率模型的藏语语音识别方法原理

(1) 语音语料库创建. 根据识别系统应用领域收集、整理文本语料形成文本语料库. 将文本语料按照特定语音规格进行录制, 再将录制的语音文件整理标注后创建语音语料库.

(2) 发音字典创建. 首先对文本语料进行音节统计, 将统计的音节按照 Wylie(威利) 转写方案进行音节拉丁转写, 其次将统计的音节通过基字丁拆分技术进行声韵母拆分, 最后由音节拉丁转写字符串和声韵母拉丁转写字符串共同创建发音字典.

(3) 特征参数提取. 主要提取 MFCC 特征, 同时根据不同方言特点融入其他特征信息.

(4) 语言建模. 语言模型的引入是为了解决字、词之间的上下文关系紊乱的问题. 简单来讲, 就是用来将识别出来的孤立字词组合成一句完整的句子. 通常创建基于 bigram 和 trigram 算法的具有上下文相关性的语言模型.

3.3 基于人工神经网络的藏语语音识别技术

由于人与动物的神经网络具有根据自然环境而自学习的能力, 所以人工神经网络的引入是为了实现语音识别程序的差异语境自适应[27]. 基于人工神经网络的语音识别原理图如图 5 所示.

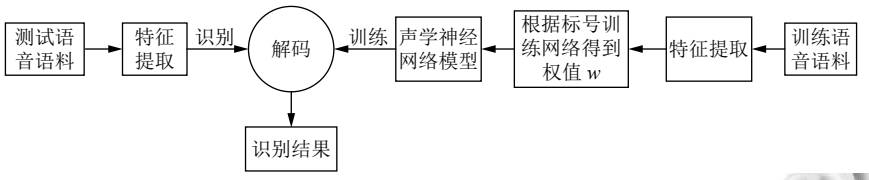

图 5 语音识别之人工神经网络原理

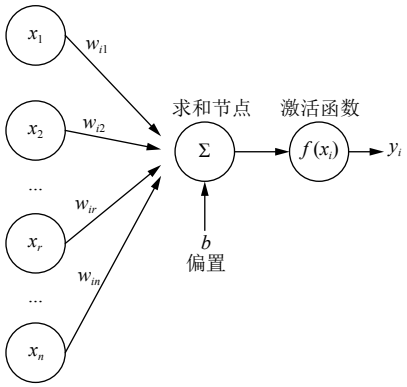

图 6 人工神经元结构

第一个要素为突触或者称为连接链集, 可以用权值来表征其大小; 第二个要素为加法器, 即线性组合器, 它将对每一个输入信号进行加权求和; 第三要素为激活函数, 它将调节信号的输出范围.

深度神经网络是人工神经网络的改进, 是更高效的网络结构. 基于深度神经网络的藏语语音识别技术是一种利用 DNN-HMM 提取深度特征并建立声学模型的藏语连续语音识别技术[17,28]. 基于深度神经网络的语音识别技术具有高抗噪能力和高识别率的特点. 基于深度神经网络的藏语语音识别方法原理图如图 7.

基于深度神经网络的藏语语音识别根据深度特征提取方式的不同又分为常用的 4 种: 基于卷积神经网络 (Convolutional Neural Networks, CNN)、深度置信网 (Deep Belief Network, DBN)、稀疏自动编码器 (Sparse Auto-Encoder, SAE) 和长短时记忆算法 (Long Short Term Memory, LSTM) 提取深度特征的藏语语音识别技术.

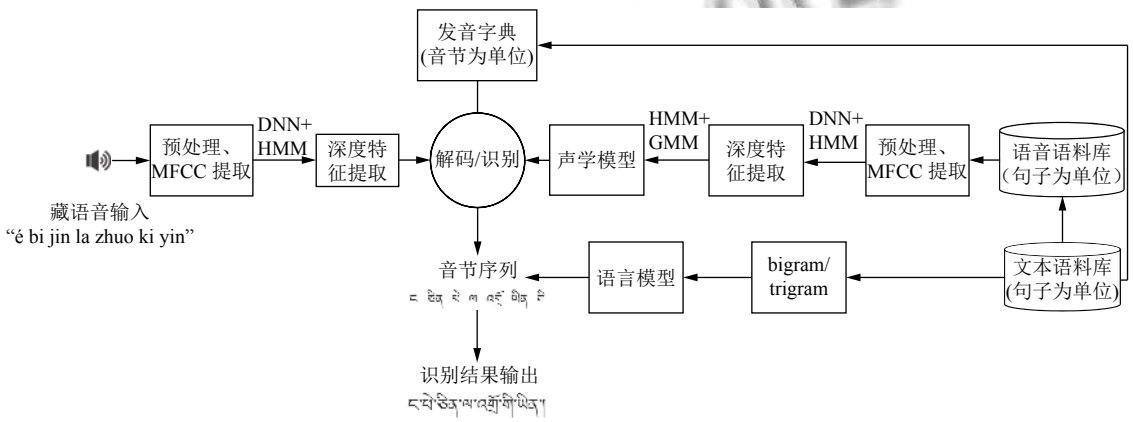

图 7 基于深度特征的藏语语音识别方法原理

3.3.1 CNN

CNN 是一种较为实用的深层神经网络的基本模型, 该模型由卷积和池化运算搭建而成[29]. 在 CNN 中, 下一层的输入是上一层的输出与某个卷积核进行卷积运算的结果, 类似依次进行逐层运算, 最终构成神经网络[26]. 基于卷积神经网络的藏语语音特征提取模型如

图 8 所示.

CNN 较其他网络模型而言, 训练需要的参数较少且具有一定的平移不变性, 是一种容易训练的模型. 在一般情况下, 只要网络结构配置的合理, 那么 CNN 建模是不需要经过预训练阶段的, 并且有时候使用随机的权值就可得到较好的特征提取结果[17].

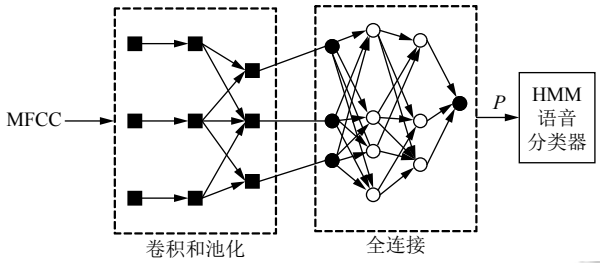

图 8 基于 CNN 的藏语语音深度特征提取模型

3.3.2 DBN

深度信念网 DBN 的方法提出是为了解决局部最优问题[16]. DBN 是由一系列受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 组成. RBM 模型如图 9 所示.

RBM 网络的特征提取层包括输入层和隐含层两层. 如图 9 所示, 其中表示观测节点的偏移量, 表示输入层与隐含层之间的权重矩阵, 表示隐含节点的偏移量. 基于 DBN 的藏语语音识别, 就是将提取的 MFCC 传统特征作为 DBN 输入, 在 DBN 的顶层用 Softmax 分类器来微调整个网络参数进而提取深度特征, 最终使用 HMM-GMM 构件模型实现藏语语音识别. 基于 DBN 的藏语语音深度特征提取模型如图 10 所示.

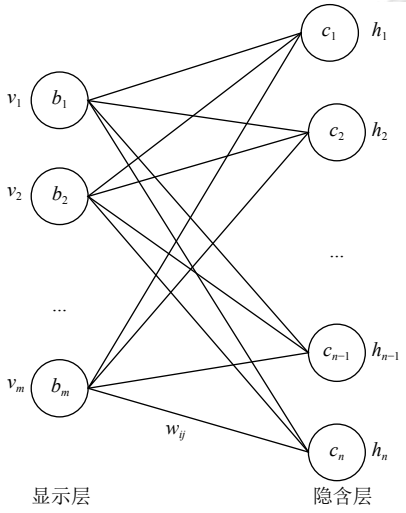

图 9 RBM 模型图

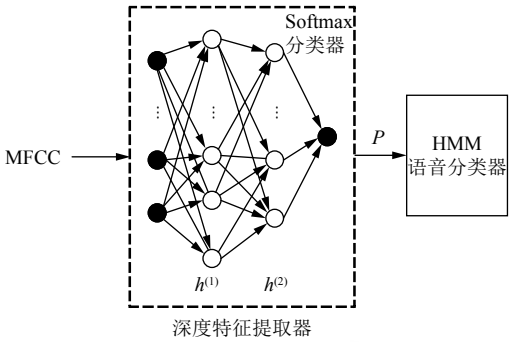

图 10 基于 DBN 的藏语语音深度特征提取模型

对于预训练阶段, 首先单独训练一个 RBM, 再把该 RBM 的输出作为下一个 RBM 的输入, 依次逐层叠加训练. 对于训练阶段, 待预训练结束后可根据误差大小使用误差传播 (Back-Propagation, BP) 算法进行有监督训练进行权值微调从而到达自动修正层次之间权值的效果, 最终形成具有深层结构的神经网络[23].

3.3.3 SAE

利用 SAE 对传统的 MFCC 声学特征进行深度学习以提取深度特征. SAE 进行特征提取的过程就是将声音频率转换为听觉神经稀疏触动信号的过程, 是一种有监督的、简单的深度特征提取方法[17]. 同 DBN 深度特征提取模型类似, 把提取的藏语语音 MFCC 特征输入到 SAE 网络中进行逐层迭代, 在网络顶层增加 Softmax 分类器来细微调整全网络的逐级参数, 最终提取得到深度学习特征. 与 DBN 不同的是 SAE 采用自下而上的逐层无监督预训练和自上而下的权重调优方式来获取语音深度特征, 这样就能成功地摆脱了参数局部最优和非稀疏性问题[29]. 基于 SAE 的藏语深度特征提取模型如图 11 所示.

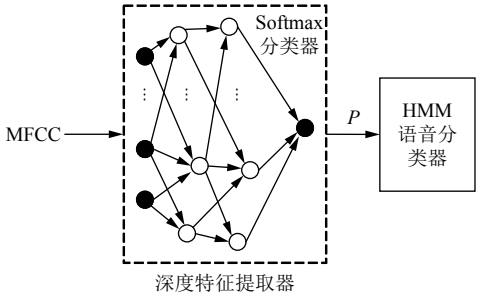

图 11 基于 SAE 的藏语深度特征提取模型

对于训练阶段, SAE 的训练过程和 DBN 训练过程一样, 使用贪心逐层预训练算法. 使用逐层贪心训练算

法训练 SAE 参数进行有监督的特征提取分为预训练和微调两步. ① 预训练时将无标签的数据样本采用无监督的方法训练网络获得参数; ② 微调时将预训练后带标签的结果数据使用 BP 算法对所有层的参数同时进行微调, 最终提取到深度特征. 在识别阶段, 同样将待识别语音经过 SAE 进行特征提取, 再将特征传入已经训练好的代表听觉中枢的 HMM 模型中进行解码, 最终实现语音识别.

3.3.4 LSTM

采用 LSTM 的输出激活与传统的 MFCC 声学特征相结合通过降维以提取深度特征[18]. 以提取的 N 维 MFCC 作为输入, 经网络迭代生成 M 维输出激活, 将此 M 维输出激活与 N 维 MFCC 特征相结合, 生成 $M+N$ 维特征, 然后使用主成分分析 (Principal Component Analysis, PCA) 算法进行降维并提取最重要的 R($N<R<M+N$) 维 Tandem 特征作为 HMM-GMM 模型的输入, 最后进行训练[25]. 基于 LSTM 的藏语特征提取流程如图 12 所示.

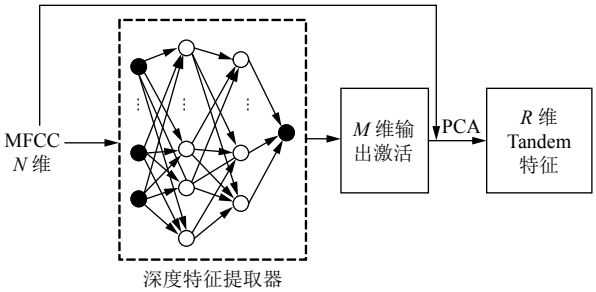

图 12 基于 LSTM 的深度特征提取过程

对于训练过程, 需经过步骤: (1) 将输入的藏语语音信号通过预处理, 再提取出 N 维 MFCC 特征; (2) 将提取的 N 为特征输入到 LSTM 网络中处理, 后输出 M 维激活 (也就是文本语料库中音节的后验概率); (3) 将 M 维输出激活与 N 维 MFCC 特征结合生成 $M+N$ 维特征, 使用 PCA 提取 R 维 Tandem 特征; (4) 将 R 维 Tandem 特征输入 HMM-GMM 模型进行训练.

对于识别过程, 首先, 需经过训练过程的步骤处理. 其次, 对照 HMM 模型库, 将最相似模型的对应文本作为输出完成识别.

3.4 技术小结

不同的藏语语音识别技术根据其特点应用到不同的场景中. 3 种技术方法各自不同的特点如表 6 所示.

基于模板匹配的藏语语音识别是简单而易实现的方法, 但是其局限在于仅适用于小词汇量孤立词、短语、短句的识别系统; 基于统计概率模型的藏语语音识别就是基于声学模型和语音模型的语音识别方法, 建模过程较为复杂, 其可适用于大词汇量的连续语音识别, 但由于其在声学建模过程中使用的每帧 MFCC 特征包含较少的语音信息量, 故抗噪声性能力弱并且易受噪声污染; 基于深度神经网络的藏语语音识别是具体高抗燥能力以及高识别效率的大词汇量连续语音识别方法, 由于深度网络的训练过程较为复杂, 导致整个方法在实现上较为困难.

表 6 3 种语音识别技术比较

技术方法	模板匹配	统计概率模型	深度神经网络
词汇量	小词汇量	大词汇量	大词汇量
识别连续性	孤立词	连续词	连续词
抗燥能力	弱	弱	强
特征参数	MFCC	MFCC	深度特征
适用范围	小	大	大
可操作性	简单	复杂	复杂

4 存在问题

藏语语音识别技术比英汉语音识别技术在研究实现上存在的困难要多得多. 目前主要存在的问题有如下这些:

(1) 在藏语表述中, 协同发音出现的情况较普遍, 故在语音端点检测时分割各语音基元 (如词、音节、音素) 间的边界比较困难, 在研究中可以考虑使用后音节对前音节元音尾作用的共振峰过渡回归方程来解决协同发音问题.

(2) 藏语虽分为卫藏、康巴、安多三大方言, 但它们各自的下属方言还比较多. 在研究中, 基本采用一些代表性的方言进行研究, 例如卫藏方言以拉萨话为代表, 安多方言以青海藏语为代表. 因此, 在研究结果应用上仍然存在较多困难. 对于这些问题, 目前只能采取“因地制宜”的研究策略.

(3) 到目前为止, 尚未有比较权威的研究用藏语语音语料库资源, 这就导致绝大多数的研究都是基于自己研究应用领域的私人语音语料进行的, 造成研究的局限以及在一定程度上阻碍了藏语语音识别的发展. 对于此问题, 寄希望于各大研究机构达成共识, 并共同创建具有代表性和研究价值的开放语音语料库.

(4) 研究的人力和财力投入不足, 也严重阻碍了藏

语语音识别的研究发展.

5 结论与展望

经过多年的研究发展, 藏语语音识别技术已经取得了良好的效果. 然而, 根据藏语各方言的发音特点, 可以在语音识别的语音去噪、端点检测方面进行技术优化来提高识别率, 其研究空间仍然很广阔. 随着深度学习软硬件资源的不断发展, 具有自学习能力、高抗噪能力和高识别率的深度学习神经网络将会成为藏语语音识别技术研究的热点和重点, 这将是未来藏语语音识别研究的趋势.

参考文献

1 拉龙东智. 藏语语音识别技术研究[硕士学位论文]. 拉萨: 西藏大学, 2015.
2 高定国, 珠杰. 藏文信息处理的原理与应用. 成都: 西南交通大学出版社, 2014. 20–21.
3 李洪波, 于洪志. 藏语语音识别的预处理研究. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议. 北京. 2006. 506–512.
4 于洪志, 李永宏, 索南楞次, 等. 安多藏语单音节声学参数数据库研究探讨. 第十一届全国民族语言文字信息学术研讨会论文集. 西双版纳. 2007. 6–11.
5 刘静萍, 德熙嘉措. 安多藏语辅音识别的设计. 民族语言文字信息技术研究——第十一届全国民族语言文字信息学术研讨会论文集. 西双版纳. 2007.
6 武光利, 戴玉刚, 马宁. 基于短时平均幅度和短时平均过零率的藏语语音端点检测研究. 福建电脑, 2007, (3): 116–122. [doi: 10.3969/j.issn.1673-2782.2007.03.067]
7 李洪波, 于洪志. 基于藏语语音学知识的语音端点检测研究. 第七届中文信息处理国际会议. 武汉. 2007. 644–649.
8 李勇, 于洪志, 达哇彭措. 基于关联规则的藏语语音韵律参数提取. 微计算机信息, 2009, 25(6): 255–257.
9 姚徐, 李永宏, 单广荣, 等. 藏语孤立词语音识别系统研究. 西北民族大学学报 (自然科学版), 2009, 30(1): 29–36, 50.
10 德庆卓玛. 基于特定人小词汇量藏语语音特征值提取的研究[硕士学位论文]. 拉萨: 西藏大学, 2010.
11 韩清华, 于洪志. 基于 HMM 的安多藏语非特定人孤立词语音识别研究. 软件导刊, 2010, 9(7): 173–175.
12 刘巧凤. 基于快速沃尔什变换的藏语音识别技术[硕士学位论文]. 成都: 西南交通大学, 2011.
13 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. [doi: 10.3969/j.issn.1000-3428.2012.05.058]
14 赵尔平, 王聪华, 党红恩, 等. 藏语孤立词语音识别技术研究. 西北师范大学学报 (自然科学版), 2015, 51(5): 50–54.
15 许彦敏. 藏语连续语音识别技术研究及系统实现[硕士学位论文]. 北京: 中央民族大学, 2015.
16 王辉, 赵悦, 刘晓凤, 等. 基于深度特征学习的藏语语音识别. 东北师范大学报 (自然科学版), 2015, 47(4): 69–73.
17 刘晓凤. 藏语语音深度特征提取及语音识别研究[硕士学位论文]. 北京: 中央民族大学, 2016.
18 张宇聪. 基于深度学习的藏语拉萨方言语音识别的研究[硕士学位论文]. 兰州: 西北师范大学, 2016.
19 周楠, 赵悦, 李要嫱, 等. 基于瓶颈特征的藏语拉萨话连续语音识别研究. 北京大学学报 (自然科学版), 2018, 54(2): 249–254.
20 赵悦, 李要嫱, 徐晓娜, 等. 临近最优主动学习的藏语语音识别方法研究. 计算机工程与应用, 2018, 54(22): 156–159, 215. [doi: 10.3778/j.issn.1002-8331.1708-0052]
21 梁宁娜, 邓彦松. 基于 DTW 的藏语语音识别系统设计. 电子技术与软件工程, 2018, (10): 135.
22 黄晓辉, 李京. 基于循环神经网络的藏语语音识别声学模型. 中文信息学报, 2018, 32(5): 49–55. [doi: 10.3969/j.issn.1003-0077.2018.05.007]
23 李涛. 基于深度神经网络的语音信号特征学习研究[硕士学位论文]. 西安:陕西师范大学, 2018.
24 周楠. 基于深度学习的藏语非特定人连续语音识别研究[硕士学位论文]. 北京: 中央民族大学, 2017.
25 吴佳欣. 基于 TANDEM 特征的藏语拉萨方言语音识别的研究[硕士学位论文]. 兰州: 西北师范大学, 2018.
26 代龙翔. 面向藏语拉萨话语音识别的语音增强方法研究[硕士学位论文]. 兰州: 西北民族大学, 2018.
27 古典. 语音识别中神经网络声学模型的说话人自适应研究[硕士学位论文]. 合肥: 中国科学技术大学, 2018.
28 梅俊杰. 基于卷积神经网络的语音识别研究[硕士学位论文]. 北京: 北京交通大学, 2017.
29 李涛, 曹辉, 郭乐乐. 深度神经网络的语音深度特征提取方法. 声学技术, 2018, 37(4): 367–370.