分类号_____ 密级_____

UDC _____ 编号 10736

西北师范大学

硕士学位论文

（专业学位）

基于深度学习的藏语安多方言语音识别的研究

研 究 生 姓 名：孙婧雯

指导教师姓名、职称：杨鸿武（教授）

实践指导教师姓名、职称：

专 业 学 位 类 别：工程硕士

专 业 学 位 领 域：电子与通信工程

专 项 计 划：

二〇二〇年六月

Research on Speech Recognition of Tibetan Amdo Dialect based on Deep Learning

A Thesis Submitted to
Northwest Normal University
in partial fulfillment of the requirement
for the degree of
Electronics and Communication Engineering
by
Jingwen Sun
Supervisor :Hongwu Yang (Professor)

May, 2020

西北师范大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本学位论文引起的法律后果完全由本人承担。

学位论文作者签名：张婧雯

导师签名：杨鸿武

签字日期：2020年 6月 17日

西北师范大学学位论文版权使用授权书

本学位论文作者完全了解西北师范大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅。本人授权西北师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，可以公开学位论文的全部或部分内容。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：张婧雯

签字日期：2020年 6月17日

摘要

在人类发展的历史长河中，语音作为人类交流必不可少的一部分，一直是国内外学者研究的重点课题。如何让计算机与人类通过“语言”交流更是热门的研究对象。随着 Siri 等许多语音识别软件的出现和智能家居的兴起，智能语音处理的应用逐渐走进人们的生活，并持续地扮演重要角色。在这个大数据的时代，拥有着对数据建模超能力的深度学习算法，已经被普及于语音识别、图像处理等模式识别领域。目前，语音识别技术针对英语、日语、德语、中文等主流国际语言识别正确率高达 99%以上。但是针对像藏语这样的民族方言研究仍处在很浅显的阶段。因此，本文主要研究提高深度学习在藏语安多方言连续语音识别上的效果。本文主要工作如下：

1.建立了一个用于藏语安多方言语音识别的大规模语音语料库。我们选取了 10000 个藏语常用句子来构建藏语安多方言语料库。我们筛选了以藏语安多方言为母语的 5 位男性说话人和 5 位女性说话人，每人录制 1000 句语音，一共录制的语料库时长为 15.6 小时。再根据发音词典对文本语料进行标注，并将语料按照 3:1 的比例分别组成训练集和测试集。

2.实现了基于深度神经网络（Deep Neural Networks，DNN）和隐马尔科夫模型（Hidden Markov Model，HMM）的藏语安多方言语音识别。我们首先对原始语音进行预处理、提取特征等操作，接下来利用相应的文本训练语言模型。然后利用训练集的语料进行大量训练，生成声学模型。最后将测试集语料输入模型，通过解码识别出词序列，字错率为 28.3%。

3.实现了基于混合端到端藏语安多方言语音识别。本文分别搭建了基于连接时态分类（connectionist temporal classification, CTC）和基于 Attention 架构的端到端藏语安多方言语音识别模型，并提出了一种基于混合 CTC/Attention 的方法来优化藏语安多方言语音识别的方法。通过调整系统的 CTC 所占权重参数来提高系统精确度，优化模型。当参数取 0.2 时，混合端到端模型的字错率最低，为 31.5%。

关键词：深度学习；语音识别；特征提取；DNN-HMM；CTC；Attention

Abstract

In the long history of human development, speech, as an essential part of human communication, has always been a key topic for scholars at home and abroad. How to make computer and human communicate through language has been a hot research object. With the emergence of many speech recognition software such as siri and the rise of intelligent home, the application of intelligent speech processing has gradually entered people's lives and continues to play an important role. In this era of big data, deep learning algorithms possess superpowers in data modeling, and have been widely used in the field of pattern recognition such as speech recognition and image processing. At present, the accuracy of speech recognition technology for Multilingual languages such as English, Japanese, German, and Chinese is more than 99%. However, the research on dialects of various languages is still in a shallow stage. Therefore, this thesis focuses on improving the effect of deep learning in continuous speech recognition of Amdo dialect. The main work of this thesis are as follows：

Firstly, the thesis completed the data preparation of Tibetan Amdo dialect. We selected 10000 Tibetan common sentences to construct the Tibetan Amdo dialect corpus. We screened five male and five female speakers whose native language is Tibetan Amdo dialect, and recorded 1000 sentences each, with a total corpus time of 15.6 hours. Then the text corpus is labeled according to the pronunciation dictionary, and the collected corpus is respectively composed into a training set and a test set according to a ratio of 3: 1.

Secondly, the speech recognition of Tibetan Amdo dialect is realized based on the deep neural networks and hidden Markov model. First, we preprocess the original speech, extract features, and then use the corresponding text to train the language models. Then, a large number of training materials are used to generate the acoustic model. At the same time, text is used to train language model. Finally, the test corpus is input into the model, and the recognized word sequence is decoded and the word error rate is 28.3%.

Finally, the speech recognition of Tibetan Amdo dialect is realized based on hybrid end-to-end architecture. An end-to-end Tibetan Amdo dialect speech recognition model based on the connection temporal classification and the attention architecture separately

are established. And a method based on hybrid CTC / attention is proposed to optimize the speech recognition of Amdo dialect. By adjusting the weight parameters of CTC, the system accuracy is improved and the model is optimized. The accuracy of the system is improved by adjusting the weight parameters of CTC to optimize the model. When the parameter is 0.2, the hybrid end-to-end model has the lowest error rate, which is 31.5%.

Keywords: Deep Learning; Speech Recognition; Feature Extraction; DNN-HMM；CTC；Attention

目录

第 1 章 引言

1.1 研究背景及意义

语音作为人类情感、想法的表达，它能高效地将人的想法传递给其他人，从而使得沟通效率变得更高。随着科技的飞速发展，人类逐渐的可以制造各式各样的机器来帮人们分担工作，同时为了使机器更加智能，人与机器交流更加“友好”，语音识别技术应运而生，使机器听懂人的话的梦想得以成真。语音识别一般来说就是在任何环境下，让机器都能够精准地把人说的话转化成文本的形式或者转换成命令的形式。现阶段，国际上的一些主流语言在语音识别方面已经有很好的效果，机器听懂不同国家语言不再是难题。然而我国幅员辽阔，民族众多，各地区方言差异大，对于不同的民族来说，他们的语言之间发音不同，甚至有的民族连文字都是不同的，这也给方言的识别带来了较大的困难，同时也一定程度上限制了我国多民族文化的交流以及发展。

我国是一个多民族国家，语言类别丰富，其中藏语不仅是我国使用人数众多、十分重要的少数民族语言，而且还是除汉语之外历史最悠久、文献最丰富的语言文明遗产[1]。所以研究藏语语音识别对促进我国多民族文化交流、民族互相发展起着非常重要的作用。然而对于藏语语音识别的研究，由于研究人员不熟悉藏语，不认识藏文，语言不通，导致我国在这项技术上还不够成熟，仍处于起步阶段。藏语一共包含三大方言，分别是卫藏方言、安多方言和康巴方言[2]。目前对藏语的研究大多数都是针对卫藏方言，对于安多方言的研究则相对较少。虽然藏语安多方言语音识别在小词汇量的孤立词识别中错误率较低[3-4]，但在连续大词汇量的连续语音识别中，因为训练数据以及语言学知识的缺乏，所以导致研究进展十分缓慢。现阶段学者们对藏语安多方言连续语音识别的研究的错误率仍旧较高[5]，这些远远不能够满足人们对安多方言语音识别系统的需求。因此，继续深入探索藏语安多方言连续语音识别技术存在重要的研究意义。

1.2 国内外语音识别的研究现状

语音识别早在上世纪 50 年代就已经开始萌芽，贝尔工作室创造了第一个语音识别系统，虽然系统仅能识别出十个英文数字，但是从此拉开了人们对语音识别研究的序幕[6]。不过真正取得实际进展的则是在 20 世纪 70 年代初期。当时科技发展迅速，恰好给语音识别搭建了相应的软硬件基础，在小词汇量、孤立词的识别

方面取得了实质性进展，使用线性预测编码以及动态时间规整技术[7]解决了对齐问题。进入 80 年代以后，研究的重点逐渐转向大量词汇、非特定人连续语音识别。在研究思路上也发生了重大的变化，隐马尔科夫模型的提出使得语音识别研究进入了一个崭新的阶段，相应的识别算法也由模版匹配技术进入到基于统计模型的技术[8]。到了 21 世纪，在基于隐马尔科夫模型 (Hidden Markov Model，HMM）的语音识别系统逐步完善的同时，机器学习算法也正在不断地发展，随着人工神经网络技术的迅速崛起，使语音识别产生了新的机遇，人工神经网络具有强大的自我学习能力[9]在语音分类上应用广泛。Hinton 在 2006 年提出了深度学习(Deep Learning, DL)算法，在当时引起了深度学习的热潮，并且伴随着计算机计算能力的提高（特别是 GPU 在计算中的使用）以及大数据的出现，世界范围内很多研究机构将深度学习网络作为了研究重点，基于 基于深度神经网络（Deep Neural Networks，DNN）和隐马尔科夫模型的框架开始成为主流的方法。语音识别的范围也逐渐从孤立词识别发展到连续语音识别、说话人识别、音色识别[10]等领域。尽管基于 DNN-HMM 的方法能够达到领先水平，但是识别过程非常繁琐。研究人员开始探索直接从语音序列到文本序列转换的端到端的方法。该方法难点在于输入的序列长度是远大于输出序列长度的，模型往往需要学习语音到文本的对齐关系。同年由 Alex Graves 便提出 CTC[11]的方法来解决序列不等长的问题。再到 2015 年百度提出的基于连接时态分类（connectionist temporal classification, CTC）的 Deep Speech 系统[12]，并进行商用才使 CTC 模型受到了广泛的关注。2019 年，百度将截断流式多级 Attention 端到端模型[13]应用到输入法产品。虽然端到端的方法相比 DNN-HMM 模型简单，但是端到端系统省略了语言模型以及文本标注，就会使系统识别率有所降低，所以，端到端的方法仍存在许多待改进的地方。

1.3 藏语语音识别的研究现状

虽然语音识别的相关技术已经得到了迅速的发展，但是我们国家在语音识别技术上的研究起步比较迟，针对藏语语音识别的研究更是寥寥无几。由于藏语语料数据难以收集，藏文发音字典以及藏文分词等方面语言学知识还不完善，所以导致藏语识别的研究更加落后。直到 2005 年，研究人员在藏语语音识别的研究中才有所突破，西北民族大学的李洪波、于洪志等人第一批提出将端点检测技术[14]应用于藏语语音识别当中。再到 2009 年，姚徐、单广荣等人提取梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCC）特征参数，形成语音模板库，采用动态时间规整技术[15]实现藏语孤立词语音识别系统。紧接着韩清华等人搭建了基于隐马尔科夫模型的藏语安多方言孤立词语音识别系统[16]。2012 年，李冠宇等人基于 HTK 实现藏语拉萨话特定人大词表连续语音识别[17]。2015 年,袁胜龙等人

采用深度神经网络[18]进行藏语语音识别，但识别正确率只达到了 43.26%。为了克服藏语语言学障碍，2018 年，王庆楠采用了端到端[19]的办法进行建模，训练以字为建模单元声学模型，但模型效果仍旧不佳。2019 年，西北师范大学的周刚也曾采用端到端的建模方法对藏语卫藏方言进行研究[20]，但由于语料不足，系统仍不够完善。虽然端到端的方法已经在国内外的语音识别系统当中取得了令人瞩目的成果，但是该方法目前仍没被非常有效的引入到藏语语音识别当中。因此，本文将继续深入研究基于深度学习的藏语安多方言语音识别系统，提高系统识别效果。

1.4 本文的研究内容及结构

本文总共六个章节，在现有的语音识别研究基础上进行探索，以藏语安多方言连续语音作为研究对象，音素作为识别基元，再结合深度学习的知识，搭建了基于深度学习的藏语安多方言语音识别系统。本文结构如下：

第 1 章绪论部分简要介绍了国内外语音识别技术的起源发展和藏语安多方言语音识别国内外的研究现状，并阐述了本文研究藏语安多方言语音识别的意义。

第 2 章主要介绍了藏语安多方言，包括了藏语基本情况、藏字的基本特点以及藏语安多方言的发音特点等。

第 3 章为藏语安多方言语料库的构建部分。首先设计了一套包含 10000 句文本语料库，接下来选取 10 位说话人对着文本进行录制，最终建立了一个用于藏语安多方言语音识别的大规模语音语料库。并且还构建了发音字典，按照字典标注了文本语料库。

第 4 章搭建了基于 DNN-HMM 模型的藏语安多方言语音识别系统。详细阐述了整个识别过程的具体环节以及各部分的工作过程。并对比了基于 HMM 的声学建模方法基于 DNN-HMM 的声学建模方法的实验结果。

第 5 章搭建了基于端到端模型的藏语安多方言语音识别系统。分别对比了基于 CTC 模型、基于 Attention 模型和基于混合 CTC /Attention 模型的藏语安多方言语音识别系统错误率，并对其结果进行对比分析。

第 6 章为总结与展望。主要概况了本文的主要研究内容和已经完成的工作，以及根据本实验的不足提出了几点未来将进行的重点研究方向。

1.5 本章小结

本章详细地介绍了语音识别的技术背景，并阐明了语音识别研究现状和藏语语音识别的研究现状，同时说明了本文研究藏语安多方言语音识别的意义所在。本章还概括了本文的研究内容和章节结构，详细阐述了本文各部分的段落安排。

第 2 章 藏语安多方言介绍

本章将详细地介绍藏语起源、分布情况以及藏语的特点，并且具体地阐述了藏字的构成、藏文的书写形式。重点说明了藏语安多方言的发音特点，并将其与藏语拉萨方言进行对比。

2.1 藏语基本情况

藏语是我国藏族同胞的母语，具有上千年的历史。藏语的起源应当追溯到吐蕃时代，藏文文字是从象雄文演变过来的。藏语属于汉藏语系中藏语支，不仅我国藏族人民使用藏语，尼泊尔，印度以及巴基斯坦等[21]地方也存在使用人群，它的使用人数已经超过 800 万人占我国少数民族人口的 70%以上。藏语主要分布在我国西藏省、青海省和四川省等五大地区[22]。它与汉语普通话存在较大差异，文字与发音皆不同。

藏语一共包含三大方言，分别是卫藏方言、安多方言和康巴方言。卫藏方言，指西藏自治区境内的前藏(拉萨等地)、后藏(日喀则等地)、山南地区和阿里地区[21]的方言。康方言，是指西藏自治区昌都地区、四川省甘孜藏族自治州、木里藏族自治县、云南省迪庆藏族自治州和青海省玉树藏族自治州范围内的方言。安多方言，指青海省大部分藏区(除玉树藏族自治州之外的)，甘肃省甘南藏族自治州、天祝藏族自治县和肃南裕固族自治县境内藏族的方言[23]。其中，卫藏方言相当于藏语的普通话，而安多方言主要流行于安多藏区，这三大方言在发音方面有着很大的不同，但它们有着一个共同点，就是他们使用着同一个藏语文本系统。安多方言相较于其他藏语方言，最大的特点是安多方言不区分声调，通常只有一个习惯调。而本文主要研究的正是安多藏区人民所使用的安多方言。

2.2 藏语文字的构成

由于藏语三大方言使用着同一个藏语文本系统，所以我们只需要研究藏字的构成方式。构成藏字基本单元是音节，而藏字是由不同藏文字母按照固定的拼接规则所组成的，因此藏字也可称为拼音文字。我们将藏文字母统称为音素。藏文是由藏字和标点符号组成的，藏文与汉文一样，遵照从左到右的书写格式，每两个藏字之间用“ • ”隔开，在句子的最后一般以单垂符“།”符号结尾，而章节段落则以双垂符“།།”结尾。

藏文字母一共 34 个，其中元音字母共有 4 个，辅音字母共有 30 个，这在藏文语法《三十颂》中有明确的记载[24]。藏文的书写是以一个藏文字母为中心，上下左右叠加字母而成。这个中心字母我们称之为“基字”，三十个辅音字母皆可作为基字。每个藏字至少由一个字母构成，最多由七个字母构成。藏字的构成方式有两种，元音分别可以加在上加字上面或下加字的下面，藏字的结构图如下图 2.1：

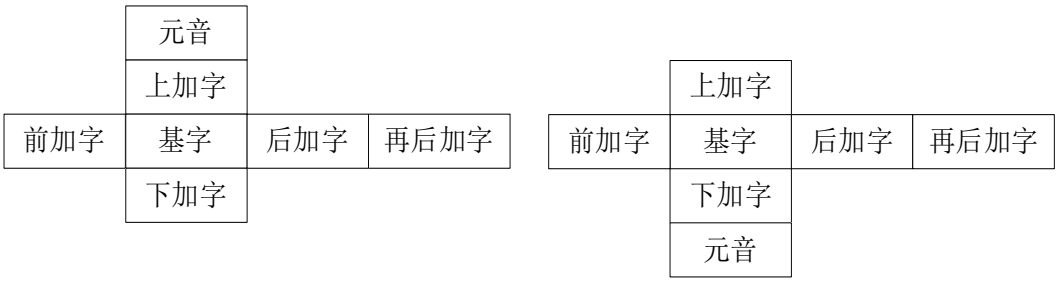

图 2.1 藏字构成方式

能够加在基字上面的元音有/i/、/e/、/o/，能够加在“下加字”下面的有元音有/u/。其他位置为辅音[25]。元音和后加字组成“韵母”其他字组成“声母”。例如图 2.2：

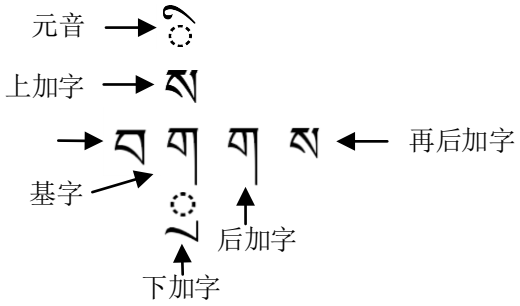

图 2.2 藏字结成图

2.3 安多方言发音特点

安多方言语作为藏语的三大方言之一，主要分散在青海牧区、甘肃甘南以及四川阿坝等地方，安多方言与青海话的词序十分相近。安多方言内部比较一致，没有土语群之分，在我国藏区的牧区中广泛通行[26]。藏语的卫藏方言和康巴方言是比较接近的，而安多方言则与他们相差巨大，安多方言保留了更多的古藏语成分，复辅音声母数量很多。安多方言的语言特性相对于另外两种方言来说，安多方言的语言特性大体概括为六个：

1.另外两种方言有复元音韵母而安多方言则没有，他们的复辅音声母少于安多

方言。

2.安多方言没有真正区分意义的声调，一般只有一种习惯调，清音读高调，浊音读低调。

3.安多方言有送气清擦音声母和浊塞擦音[27]。

4.安多方言不存在复元音和长元音。

5.安多方言很少有/c/，/ch/这样的舌面塞音。

6.安多方言使用更多单音表示动词。

我们根据藏语安多方言的发音方式，将安多方言发音总结归纳出下表 2.1：

表 2.1 藏语安多方言发音总结

	双唇音	龈音	龈音	卷舌音	龈后音	硬腭音	软腭音
鼻音	མ[m]		ན[n]		ཉ[ɲ]	ང[ŋ]	
不送气塞音	ཁ[p]		ཏ[t]			ཅ[c]	ཀ[k]
送气塞音	ཕ་བ[ph]		ཐ་དཽ[th]			ཆ་ཇ[ch]	ཁ་ག[kh]
不送气塞擦音		ཙ[ts]		ཀྵ[ts]	ཅ[tɕ]		
送气塞擦音		ཚ་ཛ[tsh]		ཁྱ་ཀྱ[tsh]	ཚ་ཛ[tɕh]		
清擦音		ས་ཟ[s]	ཤ[ɬ]	ཧ[s]	ཧ་ཅ[ɕ]		ཧ[x]
近、边、通音	ཝ[w]	ར[r]	ལ[l]			ཡ[j]	

由于藏语方言之间发音的差异，使得各个方言间声韵母的个数也各不相同。藏语安多方言和卫藏方言差别很大，其声韵母个数对照如下表 2.2：

表 2.2 安多方言和卫藏方言的声韵母对照表

	声母			韵母				
	总数	单辅音声母	复辅音声母	总数	单元音韵母	复元音韵母	带辅音韵尾的韵母	二合元音
安多方言	55	27	28	32	8	0	22	2
卫藏方言	28	28	0	17	8	2	7	0

2.4 本章小结

本章主要介绍了藏语的起源和特点，以及藏语的三大方言的区别，提出了本文重点研究的是安多方言。同时还阐述了藏语文字的构成方式以及安多方言与卫藏方言发音的特点对比。

第 3 章 藏语安多方言语料库的构建

机器训练需要大量的数据，同样的道理，为了构建完整的语音识别系统，我们需要建立完善的语料库去进行训练。目前，标准的语料库有很多，例如清华大学曾发布过的一个免费的包含 30 小时中文语音数据的中文语音数据库 THCHS-30；曾被用于小型电话识别的英语语料 TIMIT 语音数据库；为广播新闻领域的大规模连续语音识别而设计的 WSJ 语音数据库等。但是目前市面上这些常见的语料库都是根据主流国际语言去设计，而官方的藏语语料库却很少见。因为以藏语为母语的说话人不容易找到，其次目前对于藏语语言学方面的研究并不完善，很多研究者对藏文的知识不了解，导致至今藏语还没有比较完善的通用语料库，所以我们需要自己去设计藏语安多方言语料库，再进行实验。

3.1 安多方言文本语料的设计

文本语料库的设计最重要的是要科学且合理地筛选文本的内容，文本语料库的好坏直接影响到语音识别系统的训练。所以，我们在选择文本的时候，需要在保证每句话通顺且符合语法规则的同时，尽可能包含更多的发音现象，并且将每句话控制在 20 字以内。与此同时，我们还要结合藏语安多方言的发音特点以及语言学相关知识来选取语料，使其涵盖所有的有调音节、音联现象和协同发音现象。这样做的目的是缩减语料库大小并使文本更具有普遍性。为了更好的建立韵律模型，所选取的文本中要包含所有藏文的声韵母，还要包含陈述句、祈使句和复合句等各种句式。

前面章节中我们介绍过藏语方言的特点，虽然藏语的各种方言发音是不同的，但是它们采用的文本系统却是相同的。所以，我们采取从藏文的网站、藏文书籍、以及藏文日常用语中选取出 10000 个符合设计规范的藏语常用句子来构建藏语安多方言的文本语料库。

我们对每个文本进行有规律的命名，例如“ 1-fash-b”。 其中“1”代表文字数字，“fash”代表演讲者的姓名，“b”代表安多方言。此外，我们还对文本做了相应的规范化处理。例如，将每个藏文句子中的音节分隔符号“·”被空格代替，单垂符“།”被删除，仅保留基础藏字。除此之外，我们还准备了两个文本文件，一个用来存放的是所有语料的路径，另一个用来存放的是所有文本的内容集合。

3.2 安多方言语音语料的录制

我们将收集到的 10000 句藏文文本平均分为 10 组，每组共 1000 句，对每组文本选取不同说话人独立录制。我们要求说话人必须是藏族人且母语必须为藏语安多方言，同时满足吐字清晰、发音流畅、声音洪亮。需要注意的是当录音人患重感冒时存在严重的鼻音、或者因病出现声音嘶哑时，则不可参加录音。我们最终筛选了年龄都在 18 至 30 岁之间的 5 位男性和 5 位女性说话人参与录制。

录音环境的选取必须为无噪音、无回声、安静的室内。我们选取的是周围无杂声的安静的教室。录音设备需要选取带有高保真音质的话筒，我们选取的设备是智能手机。录音过程中，需要将手机设置成飞行模式，关闭网络以及其他应用程序，确保无其它干扰声音。在正式录音之前，需要对设备、人员进行测试，无误后开始正式录制。录制时，我们要求录音者在朗读前后保留 1 秒左右的静音。在录制的过程中，我们采用的是 16KHz 的采样频率以及 16bit 的量化精度录音，并将录音保存为.wav 格式。

在录制结束且校验无误后，我们需要采用 Cool Edit 语音编辑软件将收集到的语音进行手动处理，主要工作是根据波形去除语音前后大幅度的噪音。最终，我们建立的语料库总长度为 15.6 小时，我们将每个说话人录制的语料都按 4:1 的比例将其分配到测试集和训练集。其中，训练集的时长为 10.2 小时，测试集的时长为 5.4 小时。最终我们的语料库统计如下表 3.1：

表 3.1 藏语安多方言语料库统计

	说话人数（人）	句数(句)	时长（小时）	测试集(小时)	训练集(小时)
男性	5	5000	7.6	5.0	2.6
女性	5	5000	8.0	5.2	2.8
总数	10	10000	15.6	10.2	5.4

3.3 发音词典的构建

发音词典就是藏字到发音标注之间的映射。声学模型可以通过训练给出识别出来的音素，若想根据音素给出对应藏字或者句子，则必须通过语言模型和发音词典一起工作。语言模型将从发音相同、文字不同的词序列组合中找出概率最大的词序列输出。

本文选取的识别基元为音素，而藏字是由藏文字母拼接而成的，所以我们只需要将每个藏文字母标注成可识别的符号，即可将所有藏字以符号拼接的形式表示出来，进而构成藏字发音词典。

由于藏族语言文字的特殊性，不能直接被计算机所识别，所以我们需要将藏字标注成通用的形式。我国普通话的标注一般采用的是机读音标(Speech Assessment Methods Phonetic Alphabet, SAMPA)[28]直接使用键盘对语音的音标进行标注的标注方案，但是藏字不同于普通话，很多声韵母无法用现有的 SAMPA 音标符号表示出来。

因为拉丁转写是国际上通常用于表示少数民族文字的一种方式，所以本文选取的是将藏文字母转换成拉丁字母的形式标注出来，这样，藏字即可用拉丁转写的形式全部表示出来。我们制作了藏文字母到拉丁字母的转写表如表 3.2：

表 3.2 藏语安多方言字母转换表

音素	拉丁转写	音素	拉丁转写	音素	拉丁转写	音素	拉丁转写
ཨི	i	ཀ	k	ཅ	c	ཏ	t
ཨུ	u	ཁ	kh	ཆ	ch	ཐ	th
ཨེ	e	ག	g	ཇ	j	ད	d
ཨོ	o	ང	ng	ཉ	ny	ན	n
པ	p	བ	b	ཙ	ts	ཛ	dz
ཕ	ph	མ	m	ཚ	tsh	ཝ	w
ཞ	zh	ཟ	z	འ	'	ཡ	y
ར	r	ལ	l	ཤ	sh	ས	s
ཧ	h						

根据表转写规则，即能将所有藏字转写出来。构成识别系统中必不可少的发音词典。例如藏字“ཀྭ”在字典中即可以转换成“kg”的表达形式。这是本实验中一项重要的环节。

本文将语料库中的 10000 句藏语文本进行总结，提取出所有的文本，去掉重复的藏字，总共剩下 22496 个藏字，我们将每个藏字按照字母转换表标注成拉丁转写的形式，保存到一个.txt 文档中。其中，文档一共分为三列，第一列为藏字，第二列为标注的声母，第三列为标注的韵母。最终，我们生成一个包含 22496 个藏字的发音字典。发音字典的部分截图如下图 3.1：

ཀྱགས	k	ugs
ཀྱངས	k	ungs
ཀྱད	k	ud
ཀྱནད	k	und
ཀྱབ	k	ub
ཀྱབས	k	ubs
ཀྱམས	k	ums
ཀྱརད	k	urd
ཀྱལ	k	ul
ཀྱལད	k	uld
ཅེགས	k	egs

图 3.1 发音字典格式

3.4 语料库的标注

根据前面我们介绍过的发音词典，我们可以将语料库中的文本全部进行标注成拉丁转写的形式。首先，我们需要去掉所有分隔符号，将所有文本中的每个之间以空格分开。接下来，我们开始在字典中查找每个字的声韵母标注，将其字的标注写在第二行，将声韵母的标注写在第三行。需要注意的是每个标注都要以空格隔开。由于此步骤工作量巨大，我们设计了相应的程序进行统一的标注。标注格式如下图 3.2：

```
ཁྱོད་ རང་ གི་ སྐྱེན་ པ་ ཞིག་ རེད་ ཨང་
khjod ji vngravi glen pa zhig red aang
khj od j i vngr avi gl en p a zh ig r ed a ang
```

图 3.2 文本标注

图 3.2 中第一行为藏语文本，每个藏字以空格隔开。第二行为藏字的标注，字的标注以空格隔开。第三行为音素级标注，每个音素的标注以空格隔开。

3.5 本章小结

本章详细介绍了藏语安多方言语料库构建的方法，首先设计安多方言文本语料库，然后根据文本语料库进行语音语料库的录制。同时还强调了文本设计的规范以及语音录制的注意事项。最后在录制结束后对建立好的语料库进行标注。

第 4 章 基于 DNN-HMM 的藏语安多方言语音识别

一个完整的语音识别系统可以分为训练部分和识别部分。训练部分首先是对原始语音作预处理，再将处理后的信号提取特征，去训练声学模型。同时利用文本训练语言模型。识别部分将待识别的语音经过预处理、提取特征后输入训练好的模型，最后解码输出文本。一个语音识别系统通常由六部分构成：预处理、特征提取、发音字典、声学模型、语言模型和解码[10]。语音识别系统的组成如图所示：

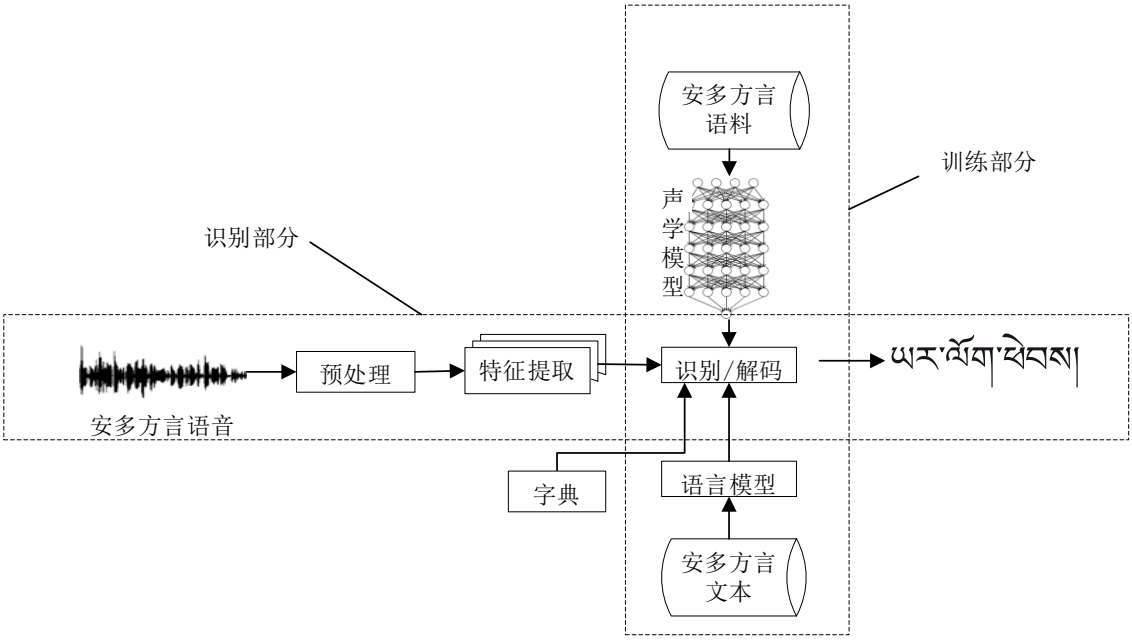

图 4.1 语音识别系统

语音识别系统通常解决的是一个从语音信号中提取到的特征向量 X 到机器输出的词序列 W 的配对问题[29]。即求解：

$$W^* = \arg\max_W P(W \mid X) \tag{4.1}$$

根据贝叶斯公式：

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)} \tag{4.2}$$

可将式(4.1)近似为：

$$W^{*}=\arg\max_{W} P(X|W)P(W) \tag{4.3}$$

这个过程中，我们用贝叶斯决策来计算后验概率，取出概率 $P(W|X)$ 的最大值，得到最可能的词序列 W^{*}[30]。其中，$P(X|W)$ 由声学模型中得到，$P(W)$ 由语言模型计算得出。

4.1 预处理

对于语音识别来说，预处理作为一个前端的技术，它的目的就是将原始的信号进行相关处理，因为语言识别系统输入的语音信号是模拟信号，我们首先需要将它进行模数转换[31]。它的能量谱在高频段和低频段的分布是不均匀的，为了更好的进行特征提取，还需要对其进行预加重处理。若想让我们的计算机知道该段音频是从什么时候开始的，又是从什么地方结束的，去除掉大量的噪音干扰就需要进行端点检测步骤。所以在进行实验之前，我们就必须对原始语音信号做出处理。通常预处理的步骤包括预加重、加窗分帧、端点检测等。预处理后该语音信号就变得更加精准，实验效果也会更好。

4.1.1 预加重

我们在进行预加重之前，需要先将模拟语音信号转换为计算机所认识数字信号，也就是说对信号模数转换。预加重是将信号里面能量低的部分进行提高，经过提高后，各频段的能量就会相对比较均匀。而我们日常生活中所采集到的语音信号通常在低频段能量很强，高频段则相对较弱。因此，我们要把高频段的能量进行加重。因为语音信号的功率谱易受口鼻以及门声激励干扰，通常干扰是当频率高达 800Hz 之后，它以 6db/Oct 进行衰减，所以高频段相对低频段来说难以求解[32]。因此，我们采用了一个数字滤波器来提高高频段的频谱：

$$H(z)=1-\alpha z^{-1} \tag{4.4}$$

式(4.4)中,α 代表预加重系数，一般取 0.9 到 1.0 之间的数值。

4.1.2 加窗分帧

在处理语音信号时，一般我们是把语音切为几个小段，这个过程我们称作分帧。常见的分帧方式有两种，连续分帧和交叠分帧。为了让帧与帧之间维持连贯性并平滑帧之间的间断，我们通常进行交叠分帧。每相邻的两帧会有重叠部分，

这部分通常被称为帧移。对语音信号分帧窗口化，会使零点左右的频谱分布的频带变得更宽，同时容易产生畸变，丢失部分能量，正因如此，我们需要加入窗函数将语音信号截断[33]。为了改善因语音信号截断而造成的频谱泄露，我们采用增加汉明窗：

$$w(n)=\begin{cases}0.54-0.46\cos[\dfrac{2\pi n}{N-1}],0\le n\le N-1\\ 0,\end{cases} \tag{4.5}$$

4.1.3 端点检测

对语音进行端点检测是指在采集到的整段音频当中，将语音和非语言的噪声分隔开的方法[34]。实际上就是指从语音音频中找到该语音部分的开始和结束，然后丢掉剩下的噪声信号部分，留下真正有效的数据，这样做不仅能够减少数据量，同时又能提高识别的质量。在端点检测这个环节，我们可以通过手工进行切音，但是由于本实验的实验数据量庞大，手工操作费时费力，所以不予采纳。因此，本文利用阈值判断来做端点检测处理。

4.2 特征提取

语音存在着许多种不同类型的特征，在训练的过程中，我们要让机器从语音信号中找出能反应这些特征的参数，从而得到特征向量，然后再利用这些特征向量进行建模训练，去预测未知语音的词序列。目前经常会使用几种特征参数包括线性预测倒谱系数（Linear Prediction Cepstral Coefficients，LPCC）、感知线性预测系数（Perceptual Linear Prediction，PLP）、MFCC 特征等[35]。在这么多的特征之间，MFCC 特征是最符合人耳听觉感知的，而且应用最广泛，因此本文只提取 MFCC 特征。

因为在时域很难观察到语音信号的特性，针对此，我们经常将其转换到频域中来获得它的能量分布。也就是说将预处理后的数字信号进行快速傅里叶变换。公式如下：

$$x_w(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi k/N},0\le k\le N \tag{4.6}$$

式(4.6)中，$x(n)$ 为预处理后的数字信号，N 为变换的点数。接下来，把频谱通过梅尔带通滤波器组进行滤波，再将频谱转换成对数形式。此滤波器组采用的是一组三角带通滤波器，它通常有 22～26 个。带通滤波器函数如下：

$$H_{m^{(k)}}=\begin{cases}\dfrac{k-f(m-1)}{f(m)-f(m-1)}, f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, f(m)\le k\le f(m+1)\\ 0,\text{其他}\end{cases} \tag{4.7}$$

式(4.7)中，m 是滤波器的编号。接着计算 $H_{m^{(k)}}$ 的对数能量：

$$（M）=\ln(\sum_{k=0}^{N-1}|X_m(k)|^2\ H_m(k)),0\le m\le M \tag{4.8}$$

经过变换后此时得到的为梅尔频率滤波特征（Mel Frequency Filter Bank，Fbank）特征。最后我们再经过离散余弦变换，进而得到 MFCC 特征:

$$C_n=\sum_{m=0}^{N-1}s(m)\cos(\frac{\pi n(m-0.5)}{M}),n=1,2,...,L \tag{4.9}$$

经过离散余弦变换之后，我们得到了 13 维的静态 MFCC 特征。最后再经过一阶、二阶差分，然后提取到 39 维 MFCC 特征参数来用于模型训练。

4.3 DNN-HMM 声学模型

语音识别系统最重要的一步就是声学模型的建立，建立的模型质量直接会影响到识别系统的质量。声学模型实现的功能就是计算 P(X|W)，也就是说给定一组词序列 W，求它的特征 X 的概率。传统的语音识别声学模型多数都是基于高斯混合模型(Gaussian Mixture Model, GMM)和 HMM 的声学模型，其中，GMM 所进行的建模是根据语音频谱的高斯分布来进行的，而 HMM 则是根据信号的时序性来进行的建模[36]。直到 20 世纪初期，机器学习的引入，使得传统的声学模型被基于神经网络的声学模型所替代，从此性能大大提高。而端到端的方法被提出后打破了传统基于深度学习的方法，并且简化了建模过程。

从结构上来说，深度神经网络其实就是 GMM 的延伸。高斯混合模型是含有一个隐含层的神经网络，它的输出层是由隐含层节点的高斯混合分量线性组合而成的。而 DNN 模型含有多个隐藏层，输入层传来的信号通过分解到隐藏层，相当于声学特征向量被转换到由隐藏层各节点所组成的新网络中，特征被转换为其他形式，虽然网络含有多个隐藏层，但每层空间的作用都是重构上一层传来的信号，最后一个隐藏层通过网络后投影到状态空间，经过多次非线性的映射，模型具备了更强的能力[37]。DNN 结构如下：

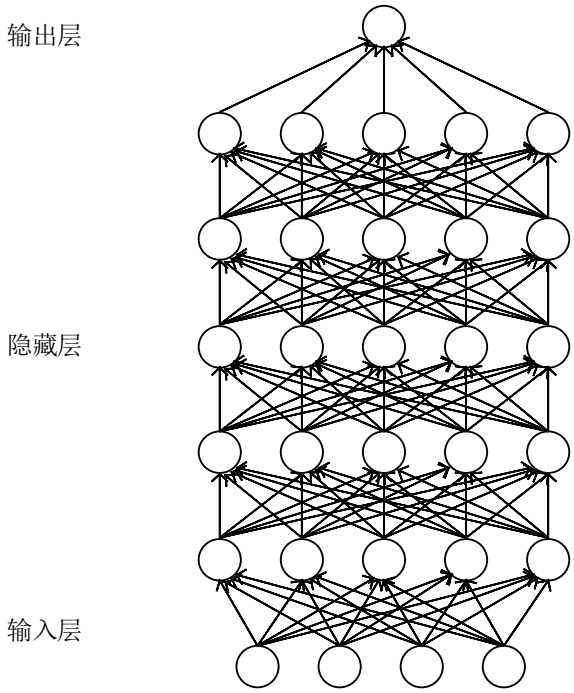

图 4.2 DNN 模型结构图

在语音识别系统中，DNN-HMM 是最常用到的模型，HMM 被用来对声音信号的序列属性进行建模，DNN 为 HMM 中的发射概率建模[38]。HMM 中强调独立性假设，即观察时刻的状态只与当前时刻有关，而与其他时刻状态无关，但实际上，相邻帧总会有一定的相关性，DNN 模型在一定程度上减弱了 HMM 的独立性假设，它的输出向量的维度等于 HMM 中状态的个数，同时每一维输出对应一个绑定的三音子状态。在训练时，我们需要在训练好的 HMM 识别系统里将训练语料强制对齐，才能得到每一帧语音在 DNN 上的目标输出值，这就意味着要想训练一个 DNN-HMM 声学模型，需先训练一个 HMM 声学模型，并通过 Viterbi 算法进行强制对齐，赋予每帧一个 HMM 状态标签，以此状态标签去训练基于深度学习的 DNN 模型[39]。最后一步，用 DNN 模型代替 HMM 模型中计算观察概率所用到的 HMM 部分，但是需要保留转移概率和初始概率等余下部分[40]。DNN-HMM 结构如下图所示：

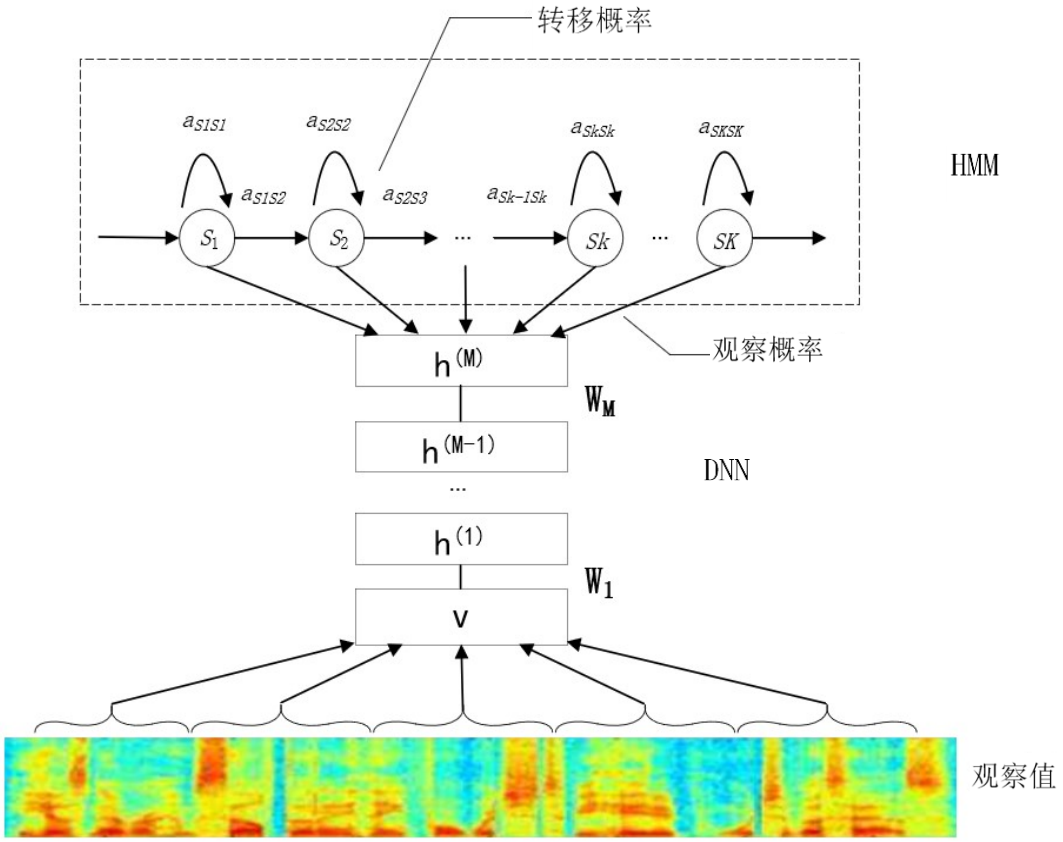

图 4.3 DNN-HMM 结构图

图中的 DNN-HMM 系统，它采用的是贝叶斯定理，同时引入 HMM 状态序列 S 的概念。所以在第三章提出的概率 $P(W \mid X)$ 可以继续进行分解：

$$W^* = \arg\max_W P(W \mid X) = \arg\max_W \sum_S P(X \mid S,W) P(S \mid W) P(W) \approx \arg\max_W \sum_S P(X \mid S) P(S \mid W) P(W) \tag{4.10}$$

在式(4.10)中采用了条件独立的假设 $P(X \mid S,W) \approx P(X \mid S)$ 简化了模型，得到的 $P(X \mid S)$ 是声学模型， $P(S \mid W)$ 是字典模型， $P(W)$ 是语言模型[41]。

声学模型 $P(X \mid S)$ 可以继续用概率链规则以及条件独立性假设来进行分解：

$$P(X \mid S) = \prod_{t=1}^{T} p(x_t \mid x_1,...,x_{t-1},S) \approx \prod_{t=1}^{T} p(x_t \mid s_t) \propto \prod_{t=1}^{T} \frac{p(s_t \mid x_t)}{p(s_t)} \tag{4.11}$$

对于字典模型 $P(S \mid W)$ 与声学模型被分解的原理一样，它也可以被分解为：

$$P(S|W)=\prod_{t=1}^{T}p(s_t|s_1,...,s_{t-1},W)\approx\prod_{t=1}^{T}p(s_t|s_{t-1},W) \tag{4.12}$$

该公式表示 HMM 状态转换器是遵照发音词典将词 W 转换成音素的一种形式。

由于我们对藏语语言学知识了解甚少，其次藏语安多方言使用人数相对较少，所能收集的语料更少，这导致 DNN-HMM 的识别性能并不是很理想。所以我们需要去寻找更加高效的模型进行语言识别。

4.4 语言模型

语音识别系统中搭建语言模型的目的就是根据声学模型输出的结果，给出对应概率最大的文字序列。语言模型是根据语言学的相关知识去建立的数学模型，它针对的训练数据不是语音，而是文字。我们通常将其分为基于统计的、基于规则的和基于神经网络的语言模型，而本文采用的是统计的 3-gram 语言模型。

N-gram 模型是假设第 N 个词的出现只与前面 N-1 个词相关，而与其它句子中出现的任何词都无关。我们可以从文本中统计 N 个词在同一个句子中同时出现的频数来计算概率。本文将 N 取 3，得到了 3-gram 语言模型 $P(W)$：

$$p(w_i|w_{i-2},w_{i-1})=\frac{count(w_{i-2},w_{i-1},w_i)}{count(w_{i-2},w_{i-1},)} \tag{4.13}$$

此时存在一个问题，若存在一个词没有被训练过，那么它的频次应为 0，词概率等于 0。但实际应用中概率不可以等于 0，因为我们的训练语料不可能会涵盖所有词组。所以，为了解决这个问题，我们设每一个词组至少出现 1 次。即不管该词组出现的频率为多少，都将频次加 1。即将式(3.34)改写成：

$$p(w_i|w_{i-2},w_{i-1})=\frac{count(w_{i-2},w_{i-1},w_i)+1}{count(w_{i-2},w_{i-1},)+1} \tag{4.14}$$

4.5 实验结果及分析

第一步我们先从准备的语料中去生成六个文件。将所有标注文件的第一行藏语提取出来然后生成词序列文本，第二行标注提取出来生成音素序列文本，将词序列文本复制，生成语音的路径文本，句子到说话人的映射文本，说话人到句子的映射文本。

准备工作做完以后，接下来我们对语音文件进行预处理，然后开始提取出 13 维的 MFCC 特征，进而根据这些特征去计算倒谱均值以及方差归一化。

接下来我们需要做的是生成语言模型。在这部分中主要是为了生成 L.fst 和

G.fst 两个文件。其中，G.fst 是一个有限状态转换机形式的藏语语言模型，L.fst 则是有限状态转换机形式藏文发音字典。接下来开始进行声学模型的训练。首先训练单音子的 HMM 模型，将它迭代 40 次后，测试单音子模型，同时建立完全的识别网络，并且输出一个有限状态转换器，最后以语言模型和测试数据为输入计算词错率（Word Error Rate, WER）。在这个过程中，我们对数据要求强制对齐，时间一般是在训练新模型前去执行。本实验采用语言模型工具 SRILM 进行训练 3-gram 语言模型。首先将语料库中所有文本语料删除所有标点符号放到一个.txt 文件中，同时对语料进行分词并用空格隔开。完成准备工作后开始统计每个词出现的频率。最后采用最大似然估计以及平滑算法利用训练集数据训练语言模型，并根据测试集数据计算其困惑度。生成的 3-gram 模型部分截图如下：

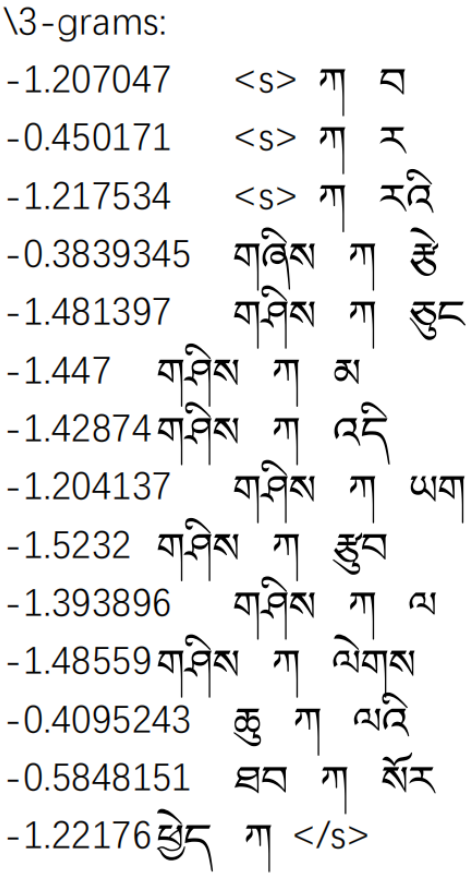

图 4.4 部分语言模型截图

我们用单音子模型作为输入训练上下文相关的三音子模型，把特征使用线性判别分析（Latent Dirichlet Allocation，LDA）和最大似然线性转换（Maximum Likelihood Linear Transformation，MLLT）进行变换之后，然后在训练中加入 LDA 以及 MLLT 的三音子模型。计算 MFCC 特征之后我们拼接了这两个算法，然后线性判别分析了 LDA 和最大似然线性转换 MLLT 进行降维，紧接着运用基于特征空

间的最大似然去线性回归来进行说话人的自适应训练，然后再对说话人自适应的模型进行解码。对于当前模型树构建之后的每个状态是基于树统计中计数的重叠判断相似性来去选择旧模型中最接近的状态。

最后我们用 GMM 模型提供的对齐进行训练 DNN 模型。并且它的帧是由每侧 5 个窗口拼接成的。特征被 LDA 转换，其中维度降至 200，之后我们采用全局均值的方法和方差归一化的方法去获得 DNN 的输入。DNN 的架构包含 4 个隐藏层，其中每个隐藏层包括 1200 个单元。基线 DNN 模型是用交叉熵标准去进行训练的，同时使用随机梯度下降算法去执行优化，批量的大小定为 256，初始学习率设置为 0.008。

基于 DNN-HMM 的藏语安多方言语音识别实验结果如下：

表 4.1 基于 DNN-HMM 的藏语安多方言语音识别结果

WER%	mono	tri1b	tri2b	tri3b	tri4b	DNN-HMM
word	51.9	47.3	32.4	31.3	29.0	28.3
phone	49.6	45.9	31.1	29.9	28.8	27.1

在此表格中，我们统计了字错率以及音素的错误率，其中错误包含替换错误、删除错误以及插入错误。其中，tri1b 表示三音子模型实验结果。tri2b 表示在 tri1 基础上进行 LDA 和 MLLT 变换的实验结果，tri3b 表示在 tri2b 基础上做了 SAT 训练的实验结果，tri4b 表示在 tri3b 基础上计算训练对齐，然后进行 2 次 FMLLR 估计迭代的实验结果。实验结果表明，在 HMM 模型下错误率是最高的，字错率为 29.0%，音素错误率则为 28.8%。在 DNN-HMM 模型下错误率是最低的，字错率为 28.3%，音素错误率则为 27.1%。

4.6 本章小结

本章分别对语音识别系统的各大组成部分进行了详细的阐述，实验部分分别搭建了基于 HMM 模型和基于 DNN-HMM 模型的藏语安多方言语音识别系统，通过对比两个系统的错误率，发现基于 DNN-HMM 模型的藏语安多方言语音识别系统效果明显较好。

第 5 章 基于端到端的藏语安多方言语音识别

本章将分别搭建基于 CTC 模型、基于 Attention 模型和基于混合 CTC /Attention 模型三种基于端到端的藏语安多方言语音识别模型。基于端到端的方法预处理和提取特征部分都和第四章介绍的基于 DNN-HMM 方法相同，不同的是基于端到端的方法跳过了基于 DNN-HMM 方法生成语言模型的步骤。

5.1 CTC 模型

端到端语音识别系统降低了使用单个网络架构构建语音识别系统的难度，它省去了传统深度学习方法中需要使用的标注、发音词典和上下文相关树，简化了复杂的建模过程[42]。正是因为这些优点在应对藏语安多方言这种低资源小语种的研究优势明显，所以本论文提出用不需要语言学知识的端到端深度学习方法来建模。

传统的 DNN-HMM 模型需要知道每帧对应的是哪个语音。而端到端方法所在意的是输出序列与输入序列是否相同，而不是预测序列和输入序列是否在某个时间点对齐[43]。

CTC 模型同样遵循贝叶斯决策。该模型使用的是一个带有一组不同字母长度为 L 的字母序列 C。在此基础上 CTC 使用空白符号“<b>”来表示每个字母边界，得到一组新的字母序列 $C'=\{<b>,c_1,<b>,c_2,<b>,...,c_L,<b>\}$，目的是解决重叠词的问题[44]。在使用空白符号后，我们把状态序列重新命名为 Z。所以后验概率 $P(C|X)$ 可以被重塑为：

$$P(C|X)=\sum_{Z}p(C|Z,X)p(Z|X)\approx\sum_{Z}p(C|Z)p(Z|X) \tag{5.1}$$

式(5.1)依旧使用条件独立假设，令 $p(C|Z,X)\approx p(C|Z)$。这样 CTC 模型就转化为去求解 CTC 声学模型 $p(Z|X)$ 和 CTC 字母模型 $p(C|Z)$。

CTC 声学模型 $p(Z|X)$ 与 DNN-HMM 的声学模型类似，我们可以进一步的用概率链规则和条件独立性假设去进行分解，从而得到：

$$P(Z|X)=\prod_{t=1}^{T}p(z_t|z_1,...,z_{t-1},X)\approx\prod_{t=1}^{T}p(z_t|X) \tag{5.2}$$

式(5.2)中，后验分布 $p(z_t|X)$ 适用于所有的输入 X，并且可以很容易的使用双向长短期记忆网络（BLSTM）对其进行建模：

$$p(z_t \mid X) = Soft\max(LinB(h_t)) \tag{5.3}$$

$$h_t = BLSTM_t(X) \tag{5.4}$$

式(5.4)使用了一个激活函数 *Softmax*，*LinB* 是一个线性隐藏层，使用可学习的矩阵和偏差向量参数将隐藏的矢量 h_t 转换为高维矢量的线性层[12]。接受完整的输入序列，并在 *t* 处输出隐藏向量。

CTC 字母模型 $p(C \mid Z)$ 可以根据贝叶斯决策、概率链规则以及条件独立性假设被继续分解：

$$\begin{aligned} P(C \mid Z) &= \frac{p(Z \mid C)p(C)}{p(Z)} = \prod_{t=1}^{T} p(z_t \mid z_1,...,z_{t-1},C)\frac{p(C)}{p(Z)} \\ &\approx \prod_{t=1}^{T} p(z_t \mid z_{t-1},C)\frac{p(C)}{p(Z)} \end{aligned} \tag{5.5}$$

式(5.5)中，$p(z_t \mid z_{t-1},C)$ 指的是状态转移概率，$p(C)$ 指的是基于字母的语言模型，$p(Z)$ 指的是状态先验概率。特别的是，CTC 具有基于字母的语言模型 $p(C)$，并且，我们还可以在解码期间通过使用字母到单词的有限状态转换器将基于单词的语言模型加入 CTC 中。

状态转移概率的取值如下：

$$p(z_t \mid z_{t-1},C) \propto \begin{cases} 1 & z_t = c_l' \quad and \quad z_{t-1} = c_l' \quad for \quad all \quad possible \quad l \\ 1 & z_t = c_l' \quad and \quad z_{t-1} = c_{l-1}' \quad for \quad all \quad possible \quad l \\ 1 & z_t = c_l' \quad and \quad z_{t-1} = c_{l-2}' \quad for \quad all \quad possible \quad even \quad l \\ 0 & otherwise \end{cases} \tag{5.6}$$

上述式(5.6)一共分为四种情况。第一种代表自我转变，第二种是状态转变，第三种情况是一种特殊状态转换，当 l 是偶数时，c'_l $c'_{l\text{-}2}$ 分别表示两个字母，而 $c'_{l\text{-}1}$ 则是前面提到过的空白字符<*b*>，这种情况下，直接跳过空白符号，完成从 $c'_{l\text{-}2}$ 到 c'_l 的转换，通过上面的转换，CTC 可以做到像 DNN-HMM 一样的单音素对齐[45]。最终，我们可以把 CTC 模型最终要求解的后验概率 $P(C \mid X)$ 化简为：

$$p(C \mid X) \approx \underbrace{\sum_{Z}\prod_{t=1}^{T} p(z_t \mid z_t,C)p(z_t \mid X)}_{=p_{ctc}(C|X)}\frac{p(C)}{p(Z)} \tag{5.7}$$

5.2 Attention 模型

与 DNN-HMM 模型和 CTC 模型所不同的是，基于 Attention 的方法不需要做条件独立假设，而是直接根据概率链规则去估计出后验概率 $P(C \mid X)$ [46]：

$$p(C|X)=\underbrace{\prod_{l=1}^{L}p(c_l|c_1,...,c_{l-1},X)}_{=p_{att}(C|X)} \tag{5.8}$$

式(5.8)中，$P_{att}(C|X)$ 是一个基于 Attention 的多目标函数。我们可以通过如下方式得到：

$$ht=Encoder(X) \tag{5.9}$$

$$att=\begin{cases} ContentAttention(q_{t-1},h_t) \\ LocationAttention(\{a_{l-1}\}_{t=1}^{T},q_{l-1},h_t) \end{cases} \tag{5.10}$$

$$r_l=\sum_{t=1}^{T}a_{lt}h_t \tag{5.11}$$

$$p(c_l|c_1,...,c_{l-1},X)=Decoder(r_l,q_{l-1},c_{l-1}) \tag{5.12}$$

上述公式（5.12）代表编码器，采用的是 BLSTM 网络对其进行建模，模型的输入是从语音序列中提取到的特征向量 X，模型的输出则是隐藏向量 h_t。公式（3.22）中的 $ContentAttention$ 函数代表着基于上下文的 Attention 机制，$LocationAtention$ 函数表示位置感知 Attention 机制[47]。公式（5.11）中的 a_{lt} 表示 Attention 的权重，公式（5.12）是解码器的网络[48]。Attention 模型的训练目标是求解近似的后验概率 $P_{att}(C|X)$。

5.3 改进的混合 CTC/Attention 模型

Attention 机制预测序列的结束标签时，不会注意所有编码的帧并可能过早地结束预测标签，甚至它会通过将重点放在与上一个标签相同的部分来预测下一个标签[49]。在这种情况下，将重复预测相同的标记序列。CTC 模型强制单调对齐，并且不允许在同一帧中出现较大的跳跃或循环，这会避免 Attention 机制过早地预测序列结束标签[50]。因此，我们结合了两种方法，帧同步由 CTC 完成，而输出标签同步则是由 Attention 去完成。

本文把基于混合 CTC/Attention 机制的端到端语音识别技术应用到了藏语安多方言语音识别中，这样做有效的利用了两种架构在训练以及解码方面的优势。输入序列 $X=\{x_t\cdots x_T\}$ 转换为高级特征 $H=\{h_t\cdots h_T\}$，Attention 解码器生成字母序列 $C=\{c_1\cdots c_L\}$。混合系统结构图如下：

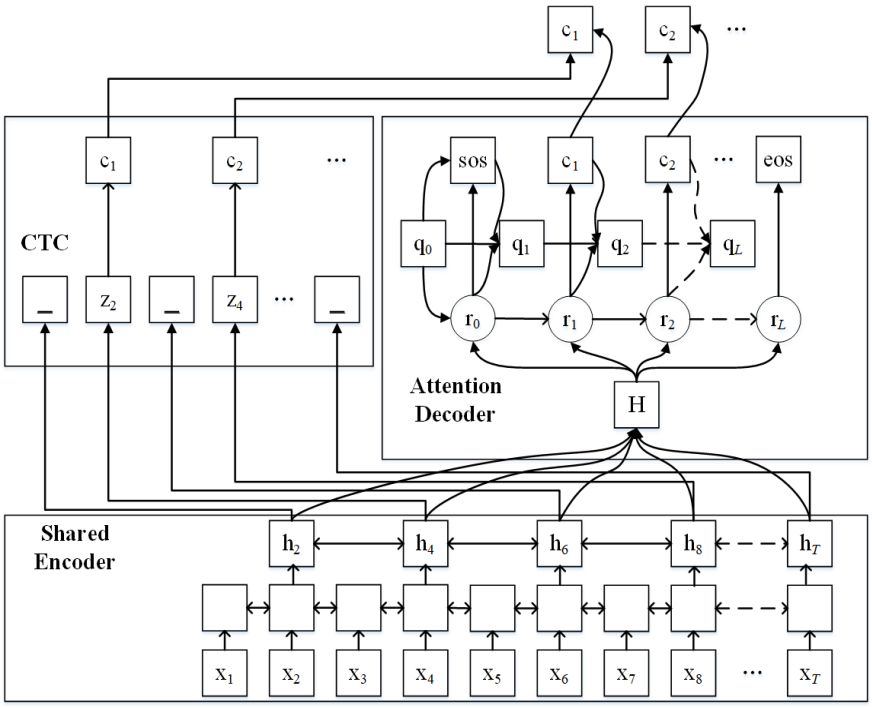

图 5.1 混合 CTC/Attention 系统结构图

编码层采用四层 BLSTM 用于训练网络，Attention 模型使用标签<sos>作为起始符号，而<eos>则代表序列的结尾。混合 CTC/Attention 的端到端普通话语音识别体系结构如图 3.4 所示。在训练过程中，我们使用多目标学习（multi-objective learning，MOL）框架，该框架结合了 CTC 和 Attention 的交叉熵来提高鲁棒性[51-52]。其中我们采用了交叉熵来用于表示损的失值：如下所示：

$$L_{MOL} = \alpha L^{CTC} + (1-\alpha)\ L^{att} \tag{5.13}$$

我们去调整多目标函数中的线性插值 α 并让它满足 $0 \le \alpha \le 1$，这样最终得到的 MOL 的交叉熵越小，则说明预测值与实际值是越接近的。当 α 等于 1 时，实验仅基于 CTC 体系结构，相反当 α 等于 0 时，实验是仅基于 Attention 架构的，当我们通过调整 α 来找出 CTC 介入多少时混合系统的效果达到最好的效果。

需要说明的是我们在混合 CTC/Attention 的系统中采用集束搜索算法来进行联合解码，使用此算法的目的是消除未对齐错误，提高准确性[53]。

5.4 实验结果及分析

我们直接将 8000 句训练集语料进行训练，将 2000 句测试集进行识别。分别测试了基于 CTC 方法、基于 Attention 方法和基于混合 CTC/Attention 方法三组实验。混合模型中我们引入了线性插值α，分别测试了α等于 0.1、0.2、0.3、0.4、0.6、0.8 时的四组不同混合模型实验用于训练和解码。实验结果如下：

表 5.1 基于混合 CTC/Attention 的藏语安多方言语音识别结果

	仅 Attention	仅 CTC	α=0.1	α=0.2	α=0.3	α=0.4	α=0.6	α=0.8
WER%	35.6	38.4	33.2	31.5	32.3	33.0	35.7	36.8
Sub%	26.5	29.2	25.2	24.1	24.7	25.4	26.6	28.0
Del%	6.9	7.9	6.6	6.2	6.4	6.3	6.5	7.2
Ins%	12.4	12.8	11.8	10.7	11.0	11.1	11.3	12.3

表格 5.1 中，WER 表示词错率，Sub 表示替换错误，Del 表示删除错误，Ins 表示插入错误。实验结果表明，在不使用包括语音词典和语言模型在内的任何语言资源的情况下，当系统仅基于 CTC 模型时，词错率达到 38.4%，当系统仅基于 Attention 模型时，词错率达到 35.6%，当 CTC 介入权重达到 0.2 时，系统错误率最低，达到 31.5%。混合 CTC /Attention 架构的收敛性、准确性和错误率分别如下图所示：

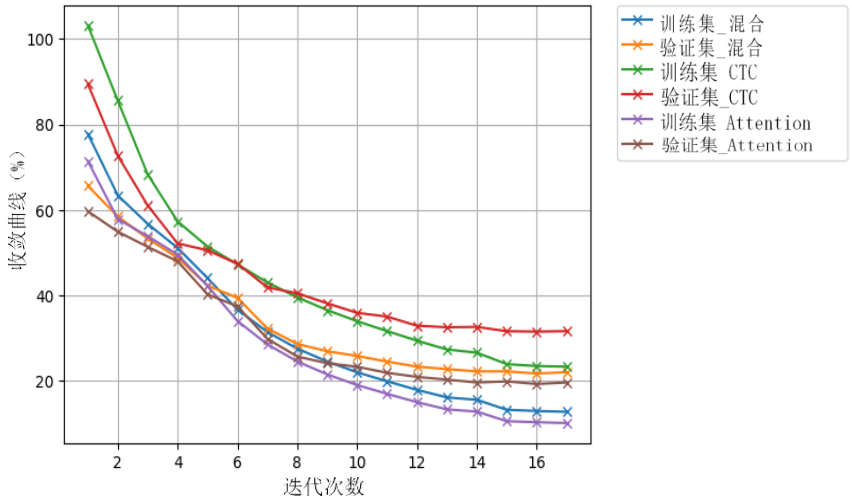

图 5.2 混合 CTC /Attention 架构的收敛曲线

图 5.2 表示混合 CTC/Attention 架构的收敛性，图中分别展示了 CTC、Attention、混合 CTC/Attention 系统训练集和验证集的损失值，横坐标为迭代次数，纵坐标为收敛程度。我们可以发现当训练迭代次数达到 15 次时，系统趋于稳定，系统的收敛性几乎不会随着训练迭代次数的扩大而产生变化。

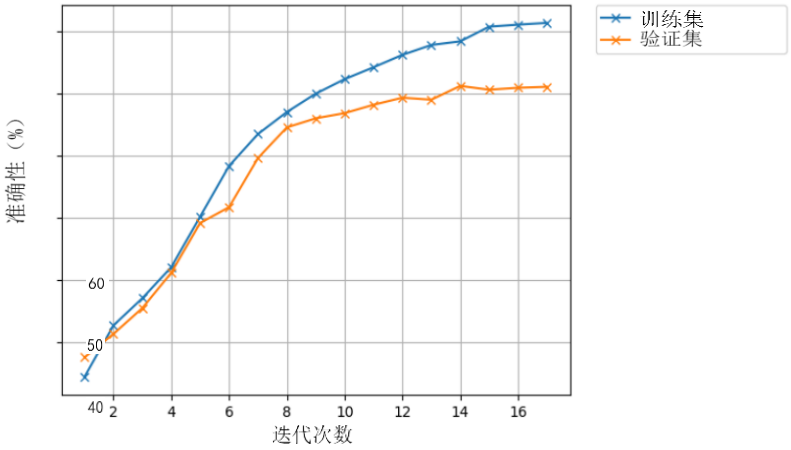

图 5.3 混合 CTC /Attention 架构的准确性

图 5.3 表示混合 CTC/Attention 架构的准确性，图中分别展示了混合 CTC/Attention 系统训练集和验证集的正确率，横坐标为迭代次数，纵坐标为正确率。我们可以发现当训练迭代次数达到 15 次时，系统趋于稳定，系统的准确性几乎不会随着训练迭代次数的扩大而产生变化。

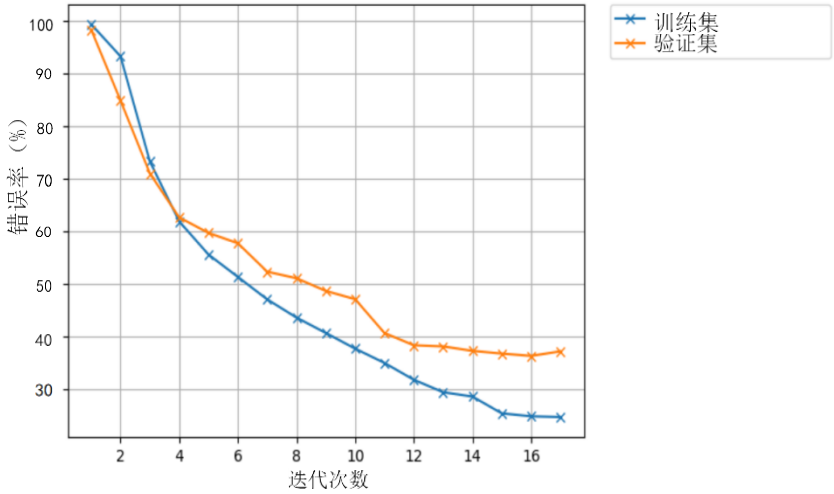

图 5.4 混合 CTC /Attention 架构的错误率

图 5.4 表示混合 CTC/Attention 架构的错误率，图中分别展示了混合 CTC/Attention 系统训练集和验证集的词错率，横坐标为迭代次数，纵坐标为错误

率。我们可以发现当训练迭代次数达到 15 次时，系统趋于稳定，系统的错误率几乎不会随着训练迭代次数的扩大而产生变化。

在相同条件下，我们构建了 10000 句藏语安多方言语料库，通过训练，进行了多个对比实验。我们分别搭建了基于 HMM 方法、基于 DNN-HMM 方法、基于 CTC 方法、基于 Attention 方法和基于混合 CTC /Attention 方法的藏语安多方言语音识别的系统，实验方法及实验结果对比如下表：

表 5.2 识别结果对比

方法	WER%
HMM	29.0
DNN-HMM	28.3
CTC	38.4
Attention	35.6
混合 CTC/Attention（α=0.2）	31.5

对比结果发现，当α取 0.2 时的混合端到端系统错误率降到最低 31.5%，比单独的仅基于 CTC 架构的错误率 38.4%和仅基于 Attention 架构的错误率 35.6%皆低，但略高于基于传统的 DNN-HMM 系统错误率 27.10%。分析原因发现，端到端系统需要大量的语料数据，本文实验数据仍旧不足。但却免去了发音字典，标注以及语言模型，更好的实现了低资源条件的端到端语音识别。

5.5 本章小结

本章分别搭建了基于 CTC 模型、基于 Attention 模型和基于混合 CTC /Attention 模型三种基于端到端的藏语安多方言语音识别系统。通过调节线性插值α，得到最优的混合模型，并将实验结果与第四章基于 HMM 模型和基于 DNN-HMM 模型的实验结果进行对比，得到基于混合 CTC /Attention 模型藏语安多方言语音识别系统错误率略高于基于传统的 DNN-HMM 系统错误率，但却免去了发音字典，标注以及语言模型，较好的实现了低资源条件的端到端语音识别。

第 6 章 总结与展望

6.1 论文总结

低资源环境下，我国针对藏语安多方言语音识别的研究不够完善。本文主要以藏语安多方言连续语音作为研究对象，音素作为研究基元，再结合深度学习的知识，搭建了多种基于深度学习的藏语安多方言语音识别系统，在省略语言模型的前提下，尽可能优化模型，将端到端识别系统的错误率逼近传统 DNN-HMM 识别系统。本文主要工作内容如下：

1. 概述了语音识别技术的起源和发展，以及藏语语音识别的研究现状，并说明了本文研究藏语安多方言语音识别的研究背景及意义。同时详细介绍了藏语安多方言的基础知识，包括藏字的基本特点以及藏字的结构等。

2. 建立了一个用于藏语安多方言语音识别的大规模语音语料库。我们构建了 10000 句的藏语安多方言语料库。我们选择了年龄都在 18 至 30 岁之间的 5 位男性和 5 位女性说话人，平均每人录制 1000 句语音，一共录制了时长为 15.6 小时的语料。再根据发音词典将文本语料进行标注到音素级，完成语料和文本对应命名。最后将语料按照 3:1 的比例分别组成训练集和测试集。

3. 完成本文对比实验部分。首先，我们对原始语音进行预处理、提取特征，利用文本训练语言模型。然后训练训练集的语料，生成基于 HMM 和基于 DNN-HMM 的藏语安多方言声学模型。最后对测试集语料进行解码，识别出词序列。接下来，为了在低资源环境下，省去传统所需的语言模型，我们分别建立基于 CTC 模型、基于 Attention 模型和基于混合 CTC/Attention 模型三种端到端藏语安多方言语音识别系统，提出了混合系统的创新，通过调整系统的 CTC 所占权重参数来提高系统精确度以此来优化模型。并将该方法与单纯基于 HMM 和基于 DNN-HMM 的方法进行了比较，得出改进后的端到端混合系统的错误率略高于传统的 DNN-HMM 系统错误率，但却免去了发音字典，标注以及语言模型。

6.2 论文展望

本文成功地在实验平台上针对藏语安多方言去搭建多种基于深度学习的语音识别系统，不过在本实验中仍然存在些许不足，需要进一步完善。经过反复的总结和反思，我将从以下几个方面的进行改进：

1.语料库的扩充

对于藏语安多方言来说，可收集到的资源十分有限。尤其是对训练一个成熟的语音识别系统来说，选取的语料越多，端到端的系统需要更多的训练数据效果才会更明显。所以我下一步工作首先需要大量扩充语料，增加实验数据。

2.克服环境噪声

因为本实验的录音设备为手机，录音环境为办公室，难免会存在一定程度的噪音，从而影响到实验质量。所以在将来，我将会选择更专业的录音设备和专门的录音棚，以此来降低噪音带来的影响。同时，在预处理部分，我打算将语音增强算法加进来使识别效果变得更好，相信一定能弥补本实验中的不足之处，降低错误率。

3.提取语音的其他特征

训练声学模型之前，需要对原始语音提取特征。语音的特征有很多，例如 PLP 特征、LPCC 特征、Tandem 特征和 Bottleneck 特征等。本文采取的是提取最常用的 MFCC 特征进行训练，未来我们可以朝着提取其他特征方向努力，例如提取语音信号 Tandem 特征或 Bottleneck 特征，甚至可以将多种特征融合，再进行降维后输入到声学模型中。

参考文献

[1] 梁宁娜, 邓彦松. 基于 DTW 的藏语语音识别系统设计[J]. 电子技术与软件工程, 2018(10): 135.

[2] 黄晓辉, 李京. 基于循环神经网络的藏语语音识别声学模型[J]. 中文信息学报, 2018, 32(05): 49-55.

[3] 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[J]. 软件导刊, 2010, 9(07): 173-175.

[4] 吴佳欣. 基于 TANDEM 特征的藏语拉萨方言语音识别的研究[D]. 西北师范大学, 2018.

[5] 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[J]. 北京大学学报(自然科学版), 2018, 54(02): 249-254.

[6] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning, 2009, 2(1): 1-127.

[7] Besacier L, Barnard E, Karpov A, et al. Automatic speech recognition for under-resourced languages: a survey[J]. Speech Communication, 2014, 56(8): 85-100.

[8] 更藏措毛. 基于深度神经网络的安多藏语语音识别[D]. 西宁: 青海师范大学, 2019：2.

[9] 李云红, 梁思程, 贾凯莉, 等.一种改进的 DNN-HMM 的语音识别方法[J]. 应用声学, 2019, 38(03): 371-377.

[10] 唐美丽, 胡琼, 马廷淮. 基于循环神经网络的语音识别研究[J]. 现代电子技术, 2019, 42(14): 152-156.

[11] 刘娟宏, 胡彧, 黄鹤宇. 端到端的深度卷积神经网络语音识别[J]. 计算机应用与软件, 2020, 37(04): 192-196.

[12] 郑晓琼, 汪晓, 江海升, 等.基于 RNN 和 WFST 译码的自动语音识别研究[J].信息技术, 2019, 43(06): 115-120.

[13] 拉龙东智. 藏语语音识别技术研究[D]. 拉萨: 西藏大学, 2015.

[14] 武光利, 于洪志, 戴玉刚. 藏语语音合成系统中语音信号的频谱转换与分析[J]. 西北民族大学学报(自然科学版), 2005, 6(59): 43-46.

[15] 姚徐,李永宏,单广荣, 等. 藏语孤立词语音识别系统[J]. 西北民族大学学报(自然科学报), 2009, 30(73): 29-36.

[16] 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[J]. 计算机工程, 2012,05: 189-191.

[17] 张高杰. 藏语安多话音色转换技术研究及其实现[D]. 兰州: 西北民族大学, 2010: 4.

[18] 袁胜龙, 郭武, 戴礼荣. 基于深层神经网络的藏语识别[J]. 模式识别与人工智能, 2015, 28(03): 209-213.

[19] 王庆楠. 基于序列记忆神经网络的藏语声学建模方法研究[D]. 中国科学技术大学, 2018.

[20] 杨鸿武, 周刚. 基于改进混合 CTC/attention 架构的端到端普通话语音识别[J]. 西北师范大学学报(自然科学版), 2019, 55(03): 48-53.

[21] 韦蕊.新中国 70 年藏语方言语音研究[J]. 西藏科技, 2019(09):72-77.

[22] 夏吾措. 藏语安多方言农区话的音系研究[J]. 西北民族大学学报(自然科学版), 2016, 37(04): 47-53.

[23] 陈小莹, 艾金勇. 安多方言—夏河话 SAMPA_AT 设计[J].智能计算机与应用, 2016,6(01):24-25+30.

[24] 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[J]. 计算机工程, 2012, 38(05): 189-191.

[25] 徐世鹏, 杨鸿武, 王海燕. 面向藏语语音合成的语音基元自动标注方法[J]. 计算机工程与应用, 2015, 51(06): 199-203.

[26] Wu Z,Yu H, Li G, et al. HMM-based Tibetan Lhasa speech synthesis system[J]. International Conference on Computer Science & Network Technology, 2014: 92-95.

[27] 龙从军, 刘汇丹, 吴健. 藏语音节标注研究[J]. 中文信息学报, 2017, 31(04): 89-93+99.

[28] Hinton G, Li D, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[29] Pranay Dighe, Afsaneh Asaei, Hervé Bourlard. On quantifying the quality of acoustic models in hybrid DNN-HMM ASR[J]. Speech Communication, 2020, 119.

[30] Hinton G, Li D, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[31] 金学骥, 叶秀清, 顾伟康. 预加重与 MMSE 结合的语音增强方法[J]. 传感技术学报, 2005(02): 300-302+306.

[32] 杨健, 李振鹏, 苏鹏. 语音分割与端点检测研究综述[J]. 计算机应用, 2020,40(01): 1-7.

[33] 黄晓辉, 李京. 基于循环神经网络的藏语语音识别声学模型[J]. 中文信息学报, 2018, 32(05): 49-55.

[34] 杨健, 李振鹏, 苏鹏. 语音分割与端点检测研究综述[J]. 计算机应用, 2020, 40(01): 1-7.

[35] Luke A, Cooke J, Luke C. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: why DNN surpasses GMMS in acoustic modeling[J]. International Symposium on Chinese Spoken Language Processing, 2012, 7196(8): 301-305.

[36] Bezoui M, Hssane A, Elmoutaouakkil A, et al. Speech Recognition of Moroccan Dialect Using Hidden Markov Models[J]. Procedia Computer Science,2019,151.

[37] Tan X, Xie Y, Ma H, et al. Corrigendum to “Recognizing the content types of network traffic based on a hybrid DNN-HMM model”[J]. Netw. Comput. Appl, 2019, 142: 51–62.

[38] 李云红,梁思程,贾凯莉, 等. 一种改进的 DNN-HMM 的语音识别方法[J]. 应用声学, 2019,38(03): 371-377.

[39] Ondřej N, Oldřich P, Ondřej G, et al. “Honza” Černocký,Lukáš Burget. Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition[J]. Computer Speech & Language,2019,58.

[40] 杨金锋, 李凯涛, 贾桂敏, 等.基于 DNN-HMM 的陆空通话声学模型构建方法[J]. 中国民航大学学报, 2019, 37(04): 36-40.

[41] 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[J]. 软件导刊, 2010, 9(07): 173-175.

[42] 蒋竺芳. 端到端自动语音识别技术研究[D]. 北京: 北京邮电大学, 2019.

[43] 刘娟宏, 胡彧, 黄鹤宇. 端到端的深度卷积神经网络语音识别[J]. 计算机应用与软件, 2020, 37(04): 192-196.

[44] Kang J, Zhang W, Liu W, et al. Lattice Based Transcription Loss for End-to-End Speech Recognition[J]. Journal of Signal Processing Systems, 2017(1): 1-11.

[45] Lee D, Lim M, Park H, et al. LSTM RNN-based Korean Speech Recognition System Using CTC[J]. Journal of Digital Contents Society, 2017, 18(1): 93-99.

[46] Miao Y, Mohammad G, and Florian M. EESEN: End-to-end speech recognition using deep rnn models and wfst-based decoding[C]//IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU), 2015: 167-174.

[47] KIM S, HORI T, WATANABE S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[R]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA, 2017, 03.

[48] 南措吉, 才让卓玛, 都格草. 基于 BLSTM 和 CTC 的藏语语音识别[J]. 青海师范大学学报(自然科学版), 2019, 35(04): 26-33.

[49] 吴威震. 基于 seq2seq 模型的聊天机器人对话研究[D]. 南京: 南京邮电大学, 2019.

[50] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Netw, 2005, 18 (5): 602.

[51] 朱小燕, 王昱, 徐伟.基于循环神经网络的语音识别模型[J]. 计算机学报, 2001, 24 (2): 213.

[52] 刘娟宏, 胡彧, 黄鹤宇. 端到端的深度卷积神经网络语音识别[J]. 计算机应用与软件, 2020,37(04): 192-196.

[53] 黄晓辉, 李京. 基于循环神经网络的藏语语音识别声学模型[J]. 中文信息学报, 2018, 32(05): 49-55.

致谢

时光如白驹过隙，我的研究生生活转瞬即逝。回顾这三年在西北师大的求学生活，可以说没有背后那些默默奉献、一直支持我的人就没有今天的我。首先我要感谢我的指导老师杨鸿武教授，他那缜密的科研态度和不苟言笑的敬业行为始终是我敬佩的标榜；他那诲人不倦的教导和平易近人的面容总能让我感受到无限温暖。能师从杨老师，我感到无比幸运，我必须对他表示真诚的感谢！

同时我也要感谢所有物理与电子工程学院的老师们，是他们对我的倾囊相授让我的学业有了大幅度提高，他们在科研的道路上严谨细致、一丝不苟的作风让我肃然起敬。在生活中对我无微不至的关怀让我感激不尽。在此我要由衷的说一声谢谢老师，你们辛苦了！

另外，我还要感谢我实验室的兄弟姐妹们，三年的实验室生活，我们一起努力，一起研究，相互监督，共同营造出我们实验室勤奋上进的气氛。我们彼此相互促进，使得每个人都在各自的研究领域收获颇丰。我很喜欢这种学习氛围，我也很庆幸和你们能够一起为实验室的整体进步贡献自己的一份力量，这不仅使我有了责任感，同时也让我变得更加优秀！

感谢我的室友，感谢你们在三年的生活中给予我的帮助。我们彼此从不同的地方来到同一个陌生的城市，是你们让我拥有如此坚固的友情，让寝室里拥有家的温馨。三年恍如昨日，毕业以后大家难以再相聚，真心祝愿你们前程似锦、工作顺利、幸福健康！

最后，我还要感谢我的爸爸、妈妈，他们是我求学生崖避风的港湾。他们一直在身后支持着我，每当我遇到不如意的事情，他们总会为我排除艰难，他们的爱一直支撑着我奋勇向前，你们对我的付出我此生都无以回报，我定不会辜负你们的期望，今后让我来为你们撑起一片天！

毕业的钟声即将敲响，论文也迎来了尾声，但我相信学术永无止境，仍有更多的知识等待着我去探索和追求。在今后的工作和生活中，我也一定会继续勇敢前行，所向披靡，争取更大的进步。

个人简历、在学位期间发表的学术论文及研究成果

个人简历

本人 1994 年 3 月 12 日出生，2016 年 6 月本科毕业于河北工业大学通信工程专业，2020 年 6 月研究生毕业于西北师范大学电子与通信工程专业。

在学位期间发表的学术论文

[1] Sun J, Zhou G, Yang H, et al. End-to-end Tibetan Ando dialect speech recognition based on hybrid CTC attention architecture[C]// Asia-Pacific Signal and Information Processing Association, 2019:628-632.（EI Accession number: 19433375，CPCI）

[2] Wang M, Qi F, Yang H, Sun J. Dongxiang speech synthesis based on statistical parameter method[C]// Asia-Pacific Signal and Information Processing Association, 2019:601-607. (EI Accession number: 19433375, CPCI)

在学位期间发表的研究成果

[1] 杨鸿武，孙婧雯. 一种可语音控制的智能车载加湿器.计算机软件著作权[P]. 专利号:201820970124.6