分类号_____ 密级_____

U D C _____ 编号 10736

西北師範大學

工 程 硕 士 学 位 论 文

基于深度学习的藏语拉萨方言语音识别的研究

研 究 生 姓 名：张宇聪

指导教师姓名、职称：杨鸿武 教授

专 业 名 称：电子与通信工程

研 究 方 向：语音信号处理

二〇一六年五月

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包括其他人已经发表或撰写过的研究成果，也不包含为获得西北师范大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：张导联　　日期：2016.6.3

关于论文使用授权的说明

本人完全了解西北师范大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

签名：张导联　导师签名：杨鸿武　日期：2016.6.2

硕士学位论文
M. D. Thesis

基于深度学习的藏语拉萨方言语音识别的研究

Research on Tibetan Lhasa dialect speech recognition based on deep learning

张宇聪

ZHANG Yucong

二○一六年五月

摘　要

让机器听懂人类的话、根据人类的命令完成工作，这是许多科研人员多年来的努力方向。近些年随着计算机计算能力的提高以及大数据的出现，深度学习算法在各个领域取得了广泛的应用。深度学习网络是一种含有多隐含层的人工神经网络，在提取特征时，具有比传统声学特征提取器更好的表达能力。许多研究已经将深度学习算法应用到了语音识别系统当中，但是这种深度学习算法目前只应用于主流语言的语音识别中，还没有引入到藏语等少数民族语言的语音识别中。本文在藏语拉萨方言语音识别中引入深度学习算法，设计了面向藏语语音识别的语料库，采用深度学习模型--长短时记忆网络模型作为藏语声学特征提取器，然后应用隐马尔可夫模型(Hidden Markov Model，HMM)进行识别。论文主要工作与创新如下：

1.建立了一个面向藏语语音识别的藏语语料库。首先设计了一个包含 51 个藏语常见字的文本语料，在对比了藏语与汉语的发音特点后，借助现有的汉语普通话标注方案 SAMPA-SC，设计了藏语拉萨方言标注方案 SAMPA-T，最后对藏语语料进行了录音和标注（4 人参与录音，每人每个字读 30 遍，共 6120 个样本）。

2.搭建了一个基于深度学习网络的藏语声学特征提取器。采用深度学习模型--长短时记忆(Long Short Term Memory，LSTM)网络模型，并将这种网络应用在藏语语音识别当中作为声学特征提取器。利用该网络输出语料库中 51 个字的后验概率，并将这 51 维输出激活与 39 维 MFCC 特征结合后经过 PCA(Principal Component Analysis)算法降维，提取最重要的 40 维 Tandem 特征，然后将这些特征输入给 HMM 进行训练与识别。

3.实现了结合长短时记忆网络与 HMM 的藏语语音识别。应用长短时记忆网络作为藏语声学特征提取器，然后应用 HMM 进行藏语识别。实验结果表明，在本文建立的语料库测试集中，本文提出的藏语语音识别方法能够达到 80.56%的识别率。

关键词：藏语语音识别；深度学习；长短时记忆网络；隐马尔可夫模型；Tandem 特征

Abstract

Researchers have been working for many years to make machines to understand human language and act as commanded. In recent years, deep learning algorithms have been widely used in various areas with the improvement of computer calculating ability and the emergence of big data. Deep learning network is a kind of artificial neural network that contains many hidden layers. Deep learning network is better than that of the traditional acoustic feature extractors in extracting acoustic features. Nowadays, many researchers have already applied deep learning algorithms in their speech recognition systems. However, the methods are only adopted in the speech recognition system for major languages. It has not been applied into minority(such as Tibetan) language speech recognition at present. Therefore the thesis introduces deep learning algorithms to Tibetan Lhasa dialect speech recognition. A Tibetan speech corpus, which including 51 of Tibetan isolate words, is designed for training corpus of speech recognition. Then a deep learning network named Long Short Term Memory network (LSTM) is used as a feature extractor to extract acoustic features from Tibetan corpus. Finally the Hidden Markov Model (HMM) is employed as a recognizer to perform speech recognition. The main works and originalities of the thesis are as follows:

Firstly, Tibetan speech corpus including 51 isolate words of Tibetan Lhasa dialect is built for Tibetan speech recognition. 51 commonly used Tibetan words were selected from text materials as a text corpus. A SAMPA-T (Tibetan) for labeling the pronunciation of Tibetan Lhasa dialect is designed by comparing the pronunciation between Tibetan Lhasa dialect and Mandarin with the aid of existing the Speech Assessment Methods Phonetic Alphabet for standard Chinese (SAMPA-SC). The speech corpus was recorded by Tibetan Lhasa speakers and labeled manually. 4 Tibetan Lhasa speakers are invited to record all the 51 isolated words. Each word is read 30 times by a speaker. Finally 6120 samples are obtained..

Secondly, a feature extractor was established based on a deep learning algorithm named long-short term memory (LSTM) network for extracting acoustic features from Tibetan speech corpus. 13-dimensional Mel frequency cepstrum coefficients (MFCC) along with their first and second difference are extracted from the speech signal to obtain a 39-dimensional feature vector. 51 output activations of the network were obtained according to the posterior probability of 51 isolate words and added to the original 39-dimensional MFCC to compose a 90-dimensional acoustic feature vector. Then the thesis applies principal component analysis (PCA) to the 90-dimensional feature vector to obtain the first 40 principal components named Tandem feature for HMM-based speech recognition.

Finally, Tibetan Speech recognition was realized by combining LSTM network and

HMM. LSTM network is used as a Tibetan acoustic feature extractor, and the HMM is used for speech recognition. Experimental results show that the proposed method can achieve a recognition rate up to 80.56% on the test set.

Key words： Tibetan speech recognition; deep learning; Long Short Term Memory network; Hidden Markov Model; Tandem Feature

目　　录

第一章 引言

1.1 研究背景及意义

让机器听懂人类的话，并根据人类的命令完成工作，这是很多科研人员多年来的努力方向[1]。近些年随着语音识别技术的发展，国内的百度、讯飞，国外的苹果、微软等诸多世界著名公司争先恐后的推出了他们的语音识别系统。科研人员预计在未来十年内，语音识别技术将会进入到生活的各个领域当中并给人类带来更多的便捷。相比于其他的沟通方式，语音是人类各种交流方式中最自然、最方便、门槛最低且效率高的表达方式。下至刚会说话的孩童，上至百岁老人，不需要再对他们进行计算机方面的培训，就可以让机器听懂他们的话，这是一件非常神奇的事情。

近些年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，深度学习算法在各个领域内取得了广泛的应用，深度学习网络是一种含有多层隐含层的人工神经网络，它能通过简单的方式来表示很多复杂的函数集合。并且这种深度学习模型在提取输入数据的特征时，具有非常惊人的表达能力和建模能力。目前深度学习算法已经被国内外许多研究机构引入到其语音信号处理系统当中[2-3]，并取得了非常不错成果。

藏族是中国不可或缺的一个民族，我国约有 450 万人在使用着藏语。目前深度学习算法已经在许多语种的语音识别系统当中取得了令人瞩目的成果，但是这项技术目前还没有被引入到藏语的语音识别当中来。本文在研究了现有的语音识别技术之后，选择并搭建了一个长短时记忆网络[4-6](Long Short Term Memory, LSTM)作为藏语声学特征提取器。这种 LSTM 结构是一种被证实了具有很好语音识别效果的深度学习结构。本项探索性研究将会对促进民族间交流起到积极作用，并为以后的藏语语音识别研究提供研究基础和借鉴。

1.2 研究现状

1.2.1 语音识别发展历史及研究现状

语音识别系统的主要任务是将输入到机器内的语音信号转换成相对应的文字。语音识别系统的主要模块如图 1.1 所示。

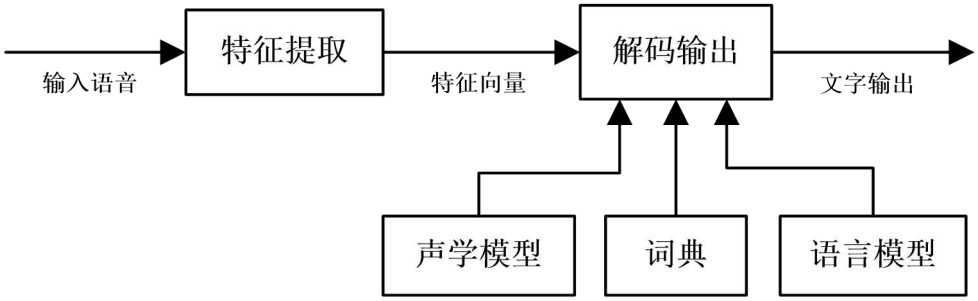

图 1.1 语音识别系统的主要模块

上世纪 50 年代，著名的贝尔实验室首先提出了一个名叫 Audrey 的数字孤立词识别系统[7]，这是最早的语音识别系统。当时主要的识别方法还是使用模拟元器件来判断共振峰的变化情况。50 年代末，英国科研人员 Denes 首次引入了统计语法这种概念并搭建了一套能够识别音素的系统[8-9]。

上世纪 60 年代，CMU 的科研人员 Raj Reddy 通过动态音素追踪的方式构建了历史上第一套连续语音识别系统[10]，这项进展为后续语音信号处理的发展做出了很大的贡献。随着动态时间规整(Dynamic Time Warping, DTW)算法[11]以及动态规整(Dynamic Programming, DP)算法的出现，科研人员解决了语音数据与模板库中的语音数据长短不一的问题。

上世纪 70 年代，线性预测编码(Linear Predictive Coding, LPC)算法[12]的提出很好的减少了语音信号所占用的空间，并且能够更好的提取输入语音的特征参数。随后 IBM 的 Fred Jelinek 将已经在其他领域略有成就的隐马尔可夫模型(Hidden Markov Model, HMM)[13]引入到了语音识别任务当中，这一研究将语音识别错误率减少了 3 倍。HMM 的引入使得语音识别系统开始向统计概率模型体系发展。

80 年代开始，科研人员逐步从孤立词语音识别任务转向了连续语音识别任务。此时，HMM 已经在语音识别任务当中取得了不错的成绩[14-16]。时至今日，HMM 依然在语音识别任务当中占据着举足轻重的地位。大词汇量连续语音识别的发展离不开 N-gram 统计语言模型[17]。80 年代中期，多层感知器(Multi-layer Perceptron, MLP) [18]的提出，使得人工神经网络(Artificial Neural Network, ANN)备受当时科研人员的欢迎，该网络通常使用反向传播算法进行训练。80 年代后期，Furui 提出的基于倒谱系数的特征参数成为了当时识别系统当中最好的语音特征参数。

90 年代，很多学者转向了声学模型训练方式的创新。主要的创新有：最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)算法 [19-20]、最大后验概率准则估计(Maximum a Posteriori Estimation, MAP)算法 [21]以及决策树状态聚类技术。此时，声学模型区分性训练(Discriminative Training, DT) [22-23]技术也有了很大的研究进展[24-25]，科研人员提出了最大互信息量准则(Maximum Mutual Information, MMI) [26-27]和最小分类错误准则(Minimum Classification Error, MCE) [28-29]。当时发布了一款 HMM 工具包(Hidden

Markov Toolkit, HTK)，这个工具包的发布很大程度上提升了国内外对语音识别的研究热度。

到了 21 世纪，在基于 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。Hinton 在 2006 年提出了深度学习(Deep Learning, DL)算法，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。深度学习网络是一种含有多层隐含层的人工神经网络结构，它可通过简单的方式来表示很多复杂的函数集合。而且在这种深度学习模型提取输入数据的特征时，具有非常惊人的表达能力和建模能力。所以目前国内外许多研究机构都已经将深度学习算法引入到了语音信号处理当中，基于 DNN-HMM 的系统与之前的 GMM-HMM 系统相比，前者将词错误率降低了大约 1/3，这项改变可以说是语音识别历史当中的一个里程碑式的标志。随着移动互联网的发展，国内外的许多知名公司相继推出了他们的语音识别 APP 来服务他们的用户，其中包括 Google 的语音助手 Google Now[30]以及 Voice Actions[31-32]、苹果公司的语音助手 Siri、Nuance 的 Dragon Go 以及 Nina、科大讯飞的语音输入法[33]。

1.2.2 藏语语音识别的研究现状

虽然现在的语音识别技术在理论方面和应用方面都取得了很好的进展，但是目前国内外对藏语的语音识别研究还相对薄弱，直到 2005 年才有相关的研究人员投入到藏语的语音识别研究当中。西北民族大学的李洪波等人是第一批从事藏语语音识别方向的科研人员[34-35]。此后，北京大学中文系[36]、中国社会科学院[37]、西北民族大学[39]、西北师范大学也都为了促进藏族信息化发展而加入到了研究队伍当中。这些科研机构的研究成果填补了国内外藏语语言信号处理的空白，并为后继的藏语语音信号处理的研究奠定了夯实的基础。

目前深度学习算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，但是该算法目前仍没被有效的引入到藏语语音识别当中。

1.3 本论文的结构

本文将一种深度学习算法—长短时记忆网络引入藏语拉萨方言语音识别当中，实现了结合长短时记忆网络与 HMM 的藏语语音识别。应用长短时记忆网络作为藏语声学特征提取器，然后应用 HMM 进行藏语识别。本文首先根据藏语拉萨方言的发音特点和国际上通用的机读音标 SAMPA 的标注方案，设计了藏语拉萨方言的机读音标 SAMPA-T。然后对 51 个藏语常见字进行了录制、切分与标注与预处理后提取出 MFCC 特征。使用 LSTM 网络对提取的 39 维的 MFCC 特征进行更近一步的处理，生成 51 个字的后验概率。将 51 维输出激活与 39 维的 MFCC 特征结合生成 90 维的 Tandem 特征，然后对 Tandem 特征使用 PCA 算法降维，取最重要的 40 维特征输入给 HMM-GMM 模型进行训练以及

识别，实现了一个基于深度学习的藏语拉萨方言的语音识别。本论文结构如下：

第 1 章，引言。首先介绍了本论文的研究背景以及将深度学习算法引入到藏语拉萨方言语音识别的研究意义。然后介绍了语音识别在发展史上的重要事件和研究现状。最后对藏语语音识别的研究现状进行了介绍。

第 2 章，藏语识别原理。重点介绍藏语语音识别系统的主要结构以及藏语语音信号处理的具体过程。包括语音信号的数字化处理、预处理和几种目前主流的特征提取方法。

第 3 章，基于 HMM 的语音识别研究。重点介绍了隐马尔可夫模型的基本原理，并介绍了 HMM 的三个基本问题和解决方法，以及公式推导。

第 4 章，深度学习神经网络模型。介绍了深度学习的历史、原理以及深度学习的 3 种常见结构。并针对长短时记忆网络进行了详细介绍和公式推导。

第 5 章，LSTM-HMM 模型的藏语语音识别实验 。首先对藏语发音和藏字进行了简单介绍，随后介绍了本文建立语料库的具体方法，并对本论文所使用的特征提取方法和 LSTM 网络配置作了进一步介绍，最后对本论文的具体实验流程和实验结果进行了说明。

第 6 章，总结与展望。对本论文实验结果进行了分析，并根据现有结果和最新的语音识别技术，提出了对未来工作的展望。

1.4 本章小结

本章主要介绍了基于深度学习的藏语拉萨方言语音识别方向的研究背景及意义。并回顾了语音识别的一些重要历史时刻和发展历程。其次介绍了语音识别系统的主要流程图，阐述了目前藏语语音识别的研究现状和目前藏语语音识别系统急需解决的问题。最后，对本论文所研究的内容和本文结构进行了阐述。

第 2 章 藏语识别原理

本章将重点介绍藏语语音识别系统的主要结构和藏语语音信号处理的具体过程。主要包括语音信号的数字化处理、预处理和当前最重要的几种特征提取方法。这些基本知识是本文研究工作的基础。

2.1 藏语语音识别系统结构

不同的语音识别任务具有不同的结构。语音识别系统一般都包括以下三个步骤：（1）语音信号的预处理；（2）经过模型训练，建立模型库；（3）输入语音与对应模型库进行匹配，输出结果。

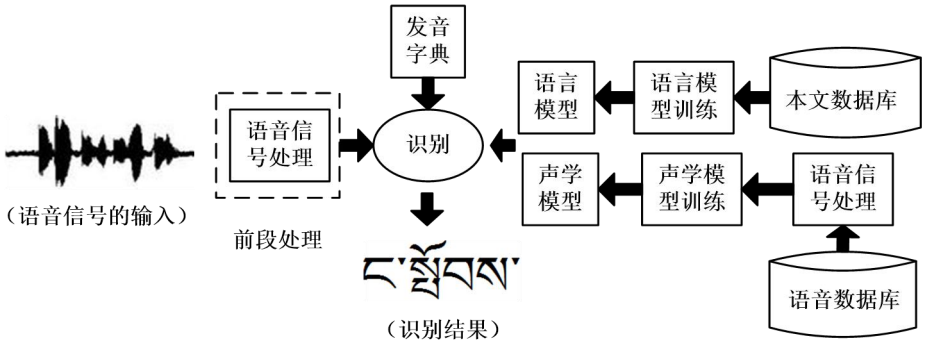

图 2.1 藏语识别系统的结构图

一般藏语拉萨方言语音识别系统的结构如图 2.1 所示，语音识别系统完成识别任务需要两个阶段：训练阶段与识别阶段。这两个阶段都需要先对输入语音进行语音信号处理。在识别阶段，系统对照训练好的声学模型进行识别。如果输入的语音是连续的话，将识别阶段的输出结果根据语音模型的句法、语法进行语音与文字的匹配，最后排除不合语法、句法的候选字，这样能大幅度的提高连续语音的识别效果。

2.2 语音信号处理与分析

语音信号是“有用信号”和“噪声（无用信号和干扰信号）”的结合，因此提取输入信号当中的有用信号是重中之重，这一过程叫做语音信号处理，它一般有 3 个过程，如图 2.2 中的虚线部分。声学模型无论是在识别阶段还是在训练阶段都需要对输入的藏语语音进行语音信号处理。语音信号处理能够提取出输入语音中的有用信息并减少存储空间。

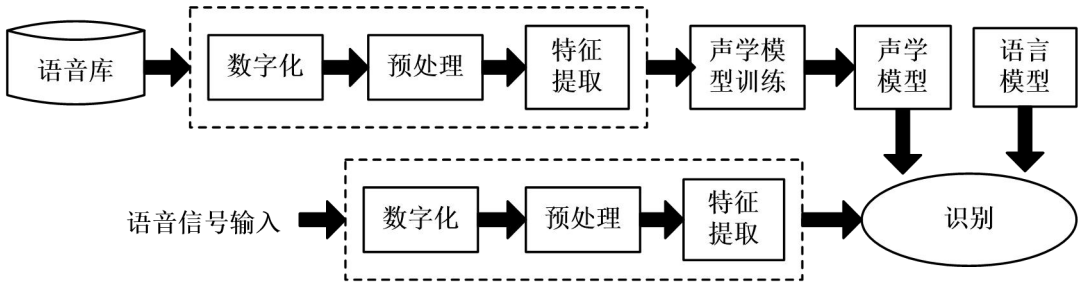

图 2.2 语音识别系统基本结构

2.2.1 语音信号的数字化处理

语音信号处理的第二步是数字化处理，经过采样器采样、量化器量化后，模拟信号变成了数字信号，如图 2.3 所示。量化能够确定信号的动态范围，采样是对信号进行等间隔抽取。

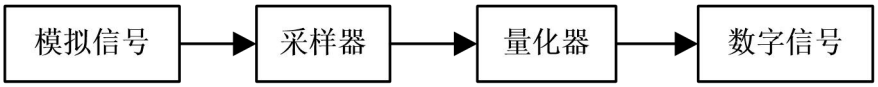

图 2.3 数字化的流程图

2.2.2 预处理

预处理对整个系统识别结果的精确度有很大的关联。语音信号的预处理主要包括如下 4 个部分：

（1）预加重

预加重是为了提高输入语音在发音过程中损失的高频部分。经过这一步骤语音信号的频谱会更加平稳，有利于后续工作。

（2）分帧

分帧是目前语音信号处理当中的一种常用方法，该方法是将一段语音信号分成许多个小段，从而就可以将一小段时间内的语音信号看作是平稳的信号，每个短时平稳的信号长度大约在 10ms 到 30ms 之间，科研人员一般会将帧与帧之间设置 1/2 的重叠，从而特征矢量系数会变得平滑。

（3）加窗

加窗的意思是把输入信号看作成许多个固定长度的窗序列 $w(n)$，然后每次都只分析 $w(n)$。这个窗函数按照时间方向移动以便分析信号。加窗运算的定义是：

$$s_n(n)=\sum_{m=-\infty}^{+\infty}s(m)w(n-m) \tag{2.1}$$

(4)端点检测

经过端点检测这一步骤，系统可以确定输入语音信号中每个基元的具体位置。目前比较常用的方法是双门限端点检测方法，这种方法的流程图如图 2.4 所示。该方法需要先给短时平均能量定一个比较小的数值，然后给过零率定一个比较大的数值。这种算法一共有四个阶段：（1）静音段；（2）过渡段；（3）语音段；（4）结束。算法的具体流程图 2.4 所示。

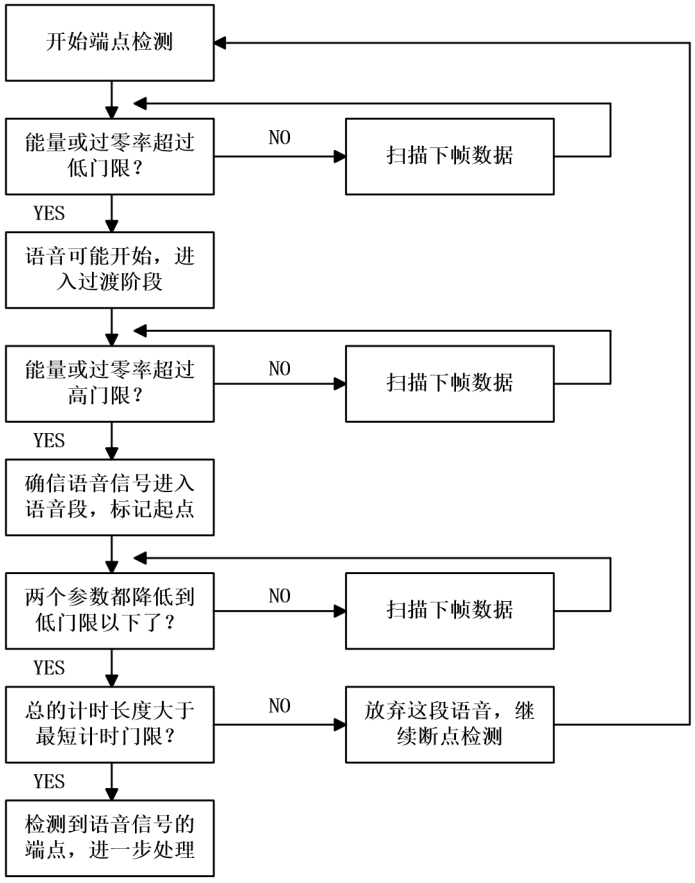

图 2.4 双门限端点检测算法的程序流程图

2.2.3 特征提取

语音信号是一种带有语音信息的信号，这种信号是有用信号和无用甚至干扰信号的结合，语音信号经过预处理过程去除干扰信号后，然后进行特征提取，这一步骤能将语音段中的有用信息提取出来。这一过程需要满足以下 4 个条件：a. 提取的特征能够代表该段语音信号；b. 提取的特征参数之间尽量独立；c. 为了保证系统实时识别和减少存储的要求，提取特征的算法应该尽可能的高效、省时；d. 特征具有区分性。常用的特征参数有以下三种：

（1）线性预测倒谱系数

线性预测倒谱系数(linear prediction cepstrum coefficient, LPCC)是复倒谱。在线性预测分析后，声道模型为：

$$H(z)=\frac{1}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{2.2}$$

p 是 LPC 线性预测器的阶数，因为：

$$H(z)=\sum_{n=1}^{\infty}h(n)z^{-n} \tag{2.3}$$

$h(n)$ 是冲激响应，$\hat{h}(n)$ 是 $h(n)$ 的复倒谱，则：

$$\hat{H}(z)=\log H(z)=\sum_{n=1}^{\infty}\hat{h}(n)z^{-n} \tag{2.4}$$

式(2.2)代入上式，等式对 z^{-1} 计算偏数，则：

$$\frac{\partial}{\partial z^{-1}}\log\left[\frac{1}{1-\sum_{k=1}^{p}a_k z^{-k}}\right]=\frac{\partial}{\partial z^{-1}}\sum_{n=1}^{\infty}\hat{h}(n)z^{-n} \tag{2.5}$$

则：

$$\frac{\sum_{k=1}^{p}ka_k z^{-k+1}}{1-\sum_{k=1}^{p}a_k z^{-k}}=\sum_{n=1}^{\infty}n\hat{h}(n)z^{-n+1} \tag{2.6}$$

所以：

$$(1-\sum_{k=1}^{p}a_k z^{-k})\sum_{n=1}^{\infty}n\hat{h}(n)z^{-n+1}=\sum_{k=1}^{p}ka_k z^{-k+1} \tag{2.7}$$

展开式(2.7)，合并同类项得：

$$\begin{aligned}
&z^{0}\left[\hat{h}(1)-a_1\right]\\
&+z^{-1}\left[2h(2)-h(1)a_1-2a_2\right]\\
&\quad\vdots\\
&+z^{-p}\left[(p+1)\hat{h}(p+1)-p\hat{h}(p)a_1-\cdots-\hat{h}(1)a_p\right]\\
&+z^{-p-1}\left[(p+2)\hat{h}(p+2)-(p+1)\hat{h}(p+1)a_1-\cdots-2\hat{h}(2)a_p\right]\\
&\quad\vdots\\
&=0
\end{aligned} \tag{2.8}$$

要让上式成立，需要将这个多项式的每项都变为 0，从而可知 $\hat{h}(n)$ 与 a_k 的推进关系，然后通过 a_k 计算 $\hat{h}(n)$：

$$\left.\begin{aligned}
&\hat{h}(0)=0 && (n\le 0)\\
&\hat{h}(1)=a_1 &&\\
&\hat{h}(n)=a_n+\sum_{k=1}^{n-1}(1-k/n)a_k\hat{h}(n-k) && (1\le n\le p)\\
&\hat{h}(n)=\sum_{k=1}^{p}(1-k/n)a_k\hat{h}(n-k) && (n>p)
\end{aligned}\right\} \tag{2.9}$$

p 为 LPC 系数的阶数，a_k 是 LPC 系数。

（2）Mel 频率倒谱系数

科研人员根据人耳的感知能力与频率的关系，提出了 Mel 频率的概念，Mel 频率与频率的关系如式(2.10)，对应关系如图 2.5 所示：

$$B(f)=1125\ln(1+f/700) \tag{2.10}$$

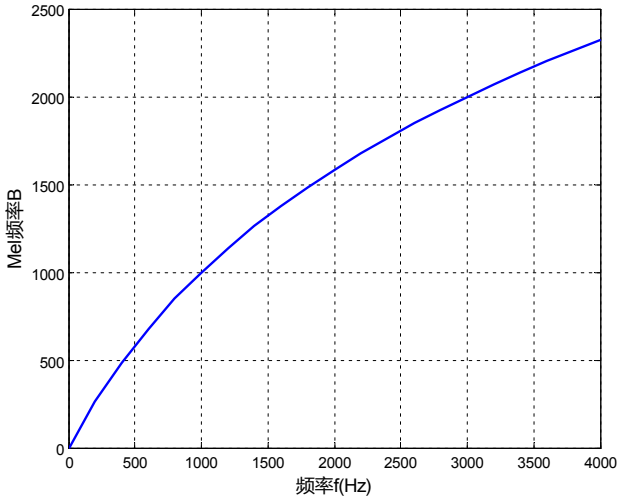

图 2.5 Mel 频率和频率的对应关系

由上述原理提出 Mel 频率倒谱系数(Mel-Frequency Cepstrum Coefficents, MFCC)，计算和提取过程如图 2.6 所示。

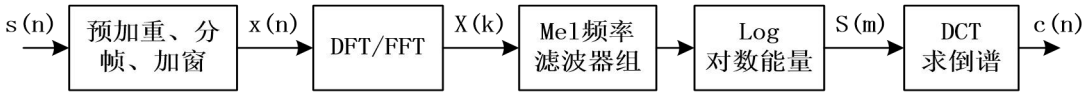

图 2.6　MFCC 的提取过程

首先，对信号 $s(n)$ 进行预处理后，$s(n)$ 变为每帧的时域信号 $x(n)$。

然后，在 $x(n)$ 后面添加若干个 0，这样就构成了一个长度为 N 的序列，然后对其进行离散傅里叶变换(Discrete Fourier Transform, DFT)，求出线性频谱 $X(k)$，$X(k)$ 的公式

为：

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N}\ ,\ 0\le n,k\le N-1 \tag{2.11}$$

为了简化系统，科研人员可能会使用快速傅立叶变换(Fast Fourier Transform, FFT)代替 DFT。

第三步，将计算所得的 $X(k)$ 通过 Mel 频率滤波器组，能够求得 Mel 频谱，然后经过对数能量处理，求得对数频谱 $S(m)$。

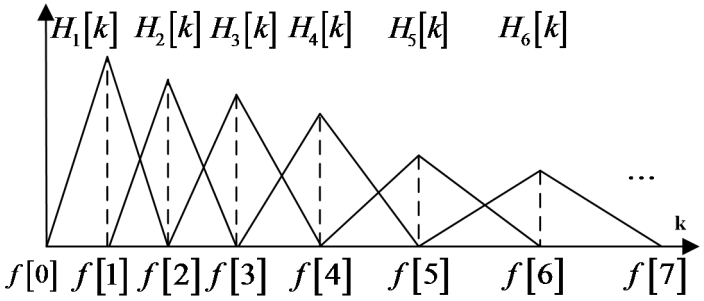

图 2.7 Mel 频率滤波器组

所有滤波器都有三角形滤波特性，中心频率为 $f(m)$，m 为 $f(m)$ 之间的距离，几个带通滤波器 $H_m(k)$ 组成一个 Mel 频率滤波器组，$0\le m<M$，M 为滤波器的数量。如图 2.7 所示。带通滤波器的传递函数为：

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m)<k\le f(m+1)\\ 0 & k>f(m+1)\end{cases},\ 0\le m<M \tag{2.12}$$

中心频率为：

$$f(m)=\left(\frac{N}{F_s}\right)B^{-1}\left(B(f_l)+m\frac{B(f_h)-B(f_l)}{M+1}\right) \tag{2.13}$$

滤波器所能应用的最高频率和最低频率分别为 f_h、f_l。N 为 DFT(FFT)的窗宽，F_s 为采样频率，B^{-1} 是 B 的逆函数：

$$B^{-1}(b)=700(e^{b/1125}-1) \tag{2.14}$$

将 Mel 频谱取对数能量会提高结果的鲁棒性。$X(k)$ 到 $S(m)$ 的总传递函数为：

$$S(m)=\ln\left(\sum_{k=0}^{N-1}\left|X(k)\right|^2H_m(k)\right),\ 0\le m<M \tag{2.15}$$

第四步， $S(m)$经过 DCT 得到 Mel 频率倒谱系数$c(n)$：

$$c(n)=\sum_{m=1}^{M-1}S(m)\cos\left(\frac{\pi n(m+1/2)}{M}\right)，\quad 0\le m<M \tag{2.16}$$

（3）基于人工神经网络的 Tandem 特征提取

近些年，随着深度学习算法的发展以及计算机硬件性能的快速提升，基于 ANN 的声学特征提取对提高语音识别系统的识别率有了令人瞩目的效果。这种基于 ANN 的声学特征叫做 Tandem 特征。在这个神经网络模型中一般会将语音信号的声学特征作为输入，使用这个声学特征参数所对应的音素作为标注。使用这些输入以及标注来对模型进行训练。模型通过训练后，对每一帧输入的特征输出一个该特征可能属于各个音素的后验概率分布。然后把这些后验概率分布作为特征传递给 HMM-GMM。

2.3 本章小结

本章首先介绍了藏语语音识别系统的主要结构，然后介绍了语音信号的数字化处理过程和预处理过程,具体介绍了预处理过程的预加重、分帧、加窗、端点检测。在介绍了特征提取应该满足的条件后,介绍了目前3种常用的语音特征参数以及它们的提取过程，这些特征分别是 LPCC、MFCC、Tandem。这些知识是本文研究工作的基础，提取好的特征对构建一个高质量的语音识别平台至关重要。

第 3 章 基于 HMM 的语音识别研究

不同的人在不同的时间说一句相同的话都是有一定区别的，有时一句话的时间长有时一句话的时间短。上世纪 60 年代，日本教授 Itakura 发明了动态时间规整(Dynamic Time Warping，DTW)算法。这种 DTW 算法的中心思想是将未知的语音信号经过动态规划来缩短或伸长信号，但是这种方法对端点的检测结果要求比较严格，不精准的端点检测使得这种算法的识别率变得很低。上世纪 70 年代，有学者将隐马尔可夫模型(Hidden Markov Model, HMM)引入到了语音识别当中，这种方法的引入使得语音识别的结果得到了很大的提高，并且一直流行至今。该模型模拟人类发音过程的随机过程，是一种针对语音识别来说很不错的统计模型。

3.1 HMM 基本原理

本文将介绍一种普通的 HMM----离散时域的有限状态自动机，通过介绍这种普通的 HMM，可以理解 HMM 其中的原理，这种状态机在每一个离散的时刻都会处于某一状态，而且其所处的所有可能状态都是已知的，然后这种状态机将会以一个特定概率向其他的已知状态跳转，如图 3.1 所示。

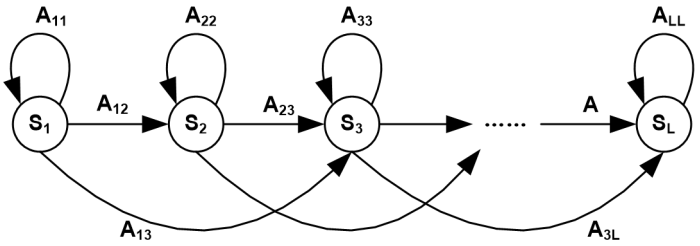

图 3.1 有限状态自动机的状态跳转示意图

因为这种自动机最开始所处的状态不一定，会通过一定的概率取到状态 $S_1 \sim S_L$ 当中的某一个，设在时间起点 $n=1$ 时，状态 $x_1 = S_l$ 的概率为 a_l，产生一个矢量 $\mathrm{a}=[a_1, a_2, \cdots, a_l, \cdots, a_L]$，该矢量被叫做初始状态概率矢量，则：

$$a_l = P(x_1 = S_l)，\ l = 1,2,\cdots,L \tag{3.1}$$

S_i 为 n 时刻的系统状态，S_j 为 $n+1$ 时刻的状态。A_{ij} 为 S_i 向 S_j 转移的概率，其中：$i, j = 1,2,\cdots,L$，所有 A_{ij} 形成的矩阵叫做状态转移概率矩阵，用 A 表示，得到：

$$A_{ij} = P(x_{n+1} = S_j \mid x_n = S_i)，\ n \geq 1，\ i, j = 1,2,\cdots,L \tag{3.2}$$

则 $\sum_{j=1}^{L} A_{ij} = 1, \forall i$ 成立，如图 3.2 所示。

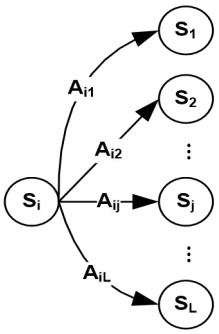

图 3.2 状态转移概率

因为状态机在$n+1$时刻只与n时刻的状态相关，与n时刻之前的所有时刻都不相关。生成的状态序列$x_1, x_2, x_3, \cdots$ 被叫做一阶马尔可夫链。

自动机在n时刻的状态x_n是未知的，只知道实R^{Q}空间的一个Q维的随机列矢量$\mathrm{y}_n = [y_{n1}, y_{n2}, \cdots, y_{nQ}]^T$，该矢量被称之为观察矢量。$\mathrm{Y} = [\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_N]$称之为观察矢量序列。如果$x_n$的状态就能确定这个观测矢量的概率，这个观测矢量就是离散分布的，所以：

$$P_{x_n=S_l}(y_n) = P(y_n \mid x_n = S_l), n \geq 1, l = 1, 2, \cdots, L \tag{3.3}$$

当y_n时，概率密度函数由x_n的状态S_l确定，有：

$$p_{x_n=S_l}(y_n) = p(y_n \mid x_n = S_l), n \geq 1, l = 1, 2, \cdots, L \tag{3.4}$$

因为上式的密度函数和概率分布只由自动机的状态决定，而不由时刻n决定，所以可以将这些密度函数以及概率分布简化成$P_{S_l}(\mathrm{y})$或$p_{S_l}(\mathrm{y})$。

将自动机每个状态的密度函数组成一个行矢量，该矢量被叫做概率密度函数矢量：$\mathrm{B} = [p_{S_1}(\mathrm{y}), p_{S_2}(\mathrm{y}), \cdots, p_{S_L}(\mathrm{y})]$。将自动机每个状态的概率分布函数叫做概率分布函数矢量，它的数学表达式为$\mathrm{B} = [P_{S_1}(\mathrm{y}), P_{S_2}(\mathrm{y}), \cdots, P_{S_L}(\mathrm{y})]$，将这两种矢量统一记做$\mathrm{B} = \{b_l(\mathrm{y})\}, l = 1, 2, \cdots, L$，可知矢量B是一个矩阵。

该自动机所处的状态是未知的，这个自动机叫做隐马尔可夫模型，通常一个HMM系统由刚才介绍的a，A，B来定义。这个HMM记为$\lambda = f(\mathrm{a}, \mathrm{A}, \mathrm{B})$

3.2 HMM 模型的三个问题

想要在语音识别中使用HMM需要解决如下几个问题：

（1）概率计算

知道了HMM的输出Y和$\lambda = f(\mathrm{a}, \mathrm{A}, \mathrm{B})$，怎样求出概率$P(\mathrm{Y} \mid \lambda)$；

（2）最优状态序列搜索：

知道了HMM的输出Y和$\lambda = f(\mathrm{a}, \mathrm{A}, \mathrm{B})$，估算出这个系统产生Y时，最应该是由哪种状态序列产生的。这第二个问题就应该算是识别的过程了。

（3）参数估计

如何根据输出Y确定一个HMM里的参数a， A， B。这里每个Y都相当于一个训练样本，一般需要很多组训练样本来训练这个HMM。这第三个问题就是训练过程。

接下来本文将对这三个问题的解决办法进行介绍。

3.2.1 概率计算

确定了参数a， A， B后，产生X并出现Y的概率是：

$$P(\mathrm{Y}\,|\,\mathrm{X},\lambda)=\prod_{n=1}^{N}P(\mathrm{y}_n\,|\,x_n,\lambda)=b_{x_1}(\mathrm{y}_1)\cdot b_{x_2}(\mathrm{y}_2)\cdots\cdots b_{x_n}(\mathrm{y}_n) \tag{3.5}$$

系统产生$\mathrm{X}=[x_1,x_2,\cdots,x_N]$的概率为：

$$P(\mathrm{X}\,|\,\lambda)=a_{x_1}\cdot A_{x_1x_2}\cdots\cdots A_{x_{N-1}x_N} \tag{3.6}$$

在确定a， A， B后，同时产生$\mathrm{X}=[x_1,x_2,;x_N]$与$\mathrm{Y}=[\mathrm{y}_1,\mathrm{y}_2,;\mathrm{y}_N]$的联合概率：

$$P(\mathrm{Y},\mathrm{X}\,|\,\lambda)=P(\mathrm{Y}\,|\,\mathrm{X},\lambda)P(\mathrm{X}\,|\,\lambda) \tag{3.7}$$

对所有可能的X求和可以得到Y的概率，则：

$$P(\mathrm{Y}\,|\,\lambda)=\sum_{all\ \mathrm{X}}P(Y\,|\,\mathrm{X},\lambda)P(\mathrm{X}\,|\,\lambda) \tag{3.8}$$

上面介绍的求概率$P(\mathrm{Y}\,|\,\lambda)$的方法在实际应用当中需要将每一个可能的状态序列相加，需要系统具有很强的计算能力。所以在实际应用中，通常会使用前向后向算法来求$P(\mathrm{Y}\,|\,\lambda)$，下面本文将介绍前向后向算法。

前向概率是用已知的$\mathrm{y}_1\mathrm{y}_2\cdots\mathrm{y}_{n-1}$概率计算$\mathrm{y}_1\mathrm{y}_2\cdots\mathrm{y}_{n-1}\mathrm{y}_n$的概率，记做$\alpha_n(j)$。从$\mathrm{y}_{n+2}\mathrm{y}_{n+3}\cdots\mathrm{y}_N$的规律求$\mathrm{y}_{n+1}\mathrm{y}_{n+2}\cdots\mathrm{y}_N$的概率叫做后向概率，记做$\beta_n(j)$。$\mathrm{Y}=[\mathrm{y}_1,\mathrm{y}_2,\cdots,\mathrm{y}_N]$的概率$P(\mathrm{Y}\,|\,\lambda)$叫做整体概率。

（1）计算$\alpha_n(j)$一般使用前向概率算法：

第1步，初始化：

$$\alpha_1(j)=a_jb_j(\mathrm{y}_1)，\ j=1,2,\cdots,L \tag{3.9}$$

第2步，递推计算：

$$\alpha_n(j)=\sum_{i=1}^{L}\alpha_{n-1}(i)A_{ij}b_j(\mathrm{y}_n)，n=2,3,\cdots,N，j=1,2,\cdots,L \tag{3.10}$$

第3步，整体概率：

$$P(\mathrm{Y}\,|\,\lambda)=\sum_{j=1}^{L}\alpha_N(j) \tag{3.11}$$

（2）计算$\beta_n(j)$用后向概率计算算法：

第1步，初始化：

$$\beta_N(j)=1，\ j=1,2,\cdots,L \tag{3.12}$$

第 2 步，递推计算：

$$\beta_n(j)=\sum_{i=1}^{L}\beta_{n+1}(i)A_{ji}b_i(\mathrm{y}_{n+1})，n=1,2,\cdots,N-1，\ j=1,2,\cdots,L \tag{3.13}$$

$n=N$ 时，将 $\alpha_n(j)=\sum_{i=1}^{L}\alpha_{n-1}(i)A_{ij}b_j(\mathrm{y}_n)$ 与 $\beta_N(j)=1$ 带入下式：

$$P(\mathrm{Y}\mid\lambda)=\sum_{j=1}^{L}\alpha_n(j)\beta_n(j) \tag{3.14}$$

所以：

$$P(\mathrm{Y}\mid\lambda)=\sum_{j=1}^{L}\alpha_N(j)\beta_N(j)=\sum_{j=1}^{L}\alpha_N(j) \tag{3.15}$$

n=1 时，将 $\alpha_1(j)=a_jb_j(\mathrm{y}_1)$ 带入 $P(\mathrm{Y}\mid\lambda)=\sum_{j=1}^{L}\alpha_n(j)\beta_n(j)$ 中，有：

$$P(\mathrm{Y}\mid\lambda)=\sum_{j=1}^{L}\alpha_1(j)\beta_1(j)=\sum_{j=1}^{L}a_jb_j(\mathrm{y}_1)\beta_1(j) \tag{3.16}$$

3.2.2 最优状态序列搜索

HMM 中，Y 对应的 X 不止一个，而且每个 X 产生 Y 的概率也不一样。所以需要由 Y 找出 X，并让 X 产生 Y 的概率最大。该算法被称为 Viterbi 算法，步骤如下：

第一步，初始化：

$$\delta_1(i)=a_ib_i(\mathrm{y}_1)，i=1,2,\cdots,L \tag{3.17}$$

第二步，递推计算：

对于 $n=1,2,\cdots,N-1$，由 $\delta_n(i)$ 求 $\delta_{n+1}(j)$，并求出 $\varphi_{n+1}(j)$：

$$\begin{aligned}&\delta_{n+1}(j)=[\max_i\{\delta_n(i)A_{ij}\}]b_j(\mathrm{y}_{n+1})\\&\phi_{n+1}(j)=\arg\max_i\{\delta_n(i)A_{ij}\},\qquad i,j=1,2,\cdots,L\end{aligned} \tag{3.18}$$

第三步，确定 δ_N：

对 $j=1,2,\cdots,L$，可以计算出 $\delta_N(j)$ 的最大值，记为 $\hat{l}_N$，则：

$$\hat{l}_N=\arg\max_j\{\delta_N(j)\} \tag{3.19}$$

第四步，路径回溯：

由 $n=N$ 开始进行回溯，得到的最优状态序列路径为：

$$\hat{l}_n=\phi_{n+1}(\hat{l}_{n+1}),\qquad n=(N-1),(N-2),\cdots,2,1 \tag{3.20}$$

3.2.3 参数估计

如何根据 HMM 的几个输出 Y 来确定 HMM 的三个参数 a， A， B 也是 HMM 需要解决的问题。这里每个 Y 都相当于一个训练样本，利用很多组训练样本训练这个 HMM，这第三个问题就是训练过程，该算法叫做 Baum-Welch 重估算法。

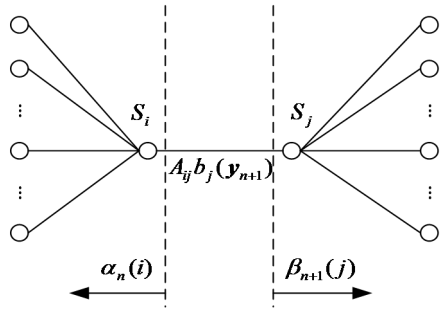

图 3.3 n 时刻向 $n+1$ 时刻，系统的状态跳转示意图

在一个 HMM 训练之前，这个模型 $\lambda = f(\mathrm{a},\mathrm{A},\mathrm{B})$，经过训练后这个模型 $\lambda' = f(\mathrm{a}',\mathrm{A}',\mathrm{B}')$，如图 3.3 所示。设 $\xi_n(i,j)$ 为这个 HMM 在 n 时刻所处于状态 S_i 然后 $n+1$ 时处在 S_j 所生成的 Y 概率，则：

$$\xi_n(i,j) = P(x_n = S_i, x_{n+1} = S_j \mid \mathrm{Y}, \lambda) \tag{3.21}$$

通过之前介绍的前向后向算法，上式为：

$$\begin{aligned}\xi_n(i,j) &= \frac{P(x_n = S_i, x_{n+1} = S_j, \mathrm{Y} \mid \lambda)}{P(\mathrm{Y} \mid \lambda)} \\ &= \frac{\alpha_n(i) A_{ij} b_j(\mathrm{y}_{n+1}) \beta_{n+1}(j)}{P(\mathrm{Y} \mid \lambda)} \\ &= \frac{\alpha_n(i) A_{ij} b_j(\mathrm{y}_{n+1}) \beta_{n+1}(j)}{\sum_{i=1}^{L} \alpha_n(i)\beta_n(i)}\end{aligned} \tag{3.22}$$

设 $\gamma_n(i)$ 为 HMM 在 n 时刻处在状态 S_i，且有 Y 的概率，所以：

$$\gamma_n(i) = P(x_n = S_i \mid \mathrm{Y}, \lambda) \tag{3.23}$$

也就是：

$$\begin{aligned}\gamma_n(i) &= \frac{P(x_n = S_i, \mathrm{Y} \mid \lambda)}{P(\mathrm{Y} \mid \lambda)} \\ &= \frac{P(x_n = S_i, \mathrm{Y} \mid \lambda)}{\sum_{n=1}^{N} P(x_n = S_i, \mathrm{Y} \mid \lambda)} \\ &= \frac{\alpha_n(i)\beta_n(i)}{\sum_{i=1}^{L} \alpha_n(i)\beta_n(i)}\end{aligned} \tag{3.24}$$

因为：

$$\gamma_n(i) = \sum_{j=1}^{L} \xi_n(i,j) \tag{3.25}$$

把 $\xi_n(i,j)$ 的所有时刻 n 相加，记做 $P\left(\mathrm{Y} \mid S_i \to S_j \ \text{ all } \ n\right)$，所以：

$$\begin{aligned}P\left(\mathrm{Y} \mid S_i \to S_j \ \text{ all } \ n\right) &= \sum_{n=1}^{N-1} \xi_n(i,j) \\ &= \frac{\sum_{n=1}^{N-1} \alpha_n(i) A_{ij} b_j(\mathrm{y}_{n+1}) \beta_{n+1}(j)}{P(\mathrm{Y} \mid \lambda)}\end{aligned} \tag{3.26}$$

然后把 $\gamma_n(i)$ $\xi_n(i,j)$ 的所有时刻 n 相加，记做 $P\left(\mathrm{Y} \mid S_i \to \text{all } S_j\right)$，得：

$$\begin{aligned}P\left(\mathrm{Y} \mid S_i \to \text{all } S_j\right) &= \sum_{n=1}^{N-1} \gamma_n(i) \\ &= \frac{\sum_{\mathrm{n}=1}^{N-1} \sum_{j=1}^{L} \alpha_n(i) A_{ij} b_j(\mathrm{y}_{n+1}) \beta_{n+1}(j)}{P(\mathrm{Y} \mid \lambda)} \\ &= \frac{\sum_{\mathrm{n}=1}^{N-1} \alpha_n(i)\beta_n(i)}{P(\mathrm{Y} \mid \lambda)}\end{aligned} \tag{3.27}$$

从 S_i 向 S_j 转移的概率的新估计为 A'_{ij}，则：

$$\begin{aligned}A'_{ij} &= \frac{P\left(\mathrm{Y} \mid S_i \to S_j \ \text{all} \, n\right)}{P\left(\mathrm{Y} \mid S_i \to \text{all } S_j\right)} \\ &= \frac{\sum_{n=1}^{N-1} \xi_n(i,j)}{\sum_{n=1}^{N-1} \gamma_n(i)}\end{aligned} \tag{3.28}$$

所以：

$$A'_{ij}=\frac{\sum_{n=1}^{N-1}\alpha_n(i)A_{ij}b_j(\mathrm{y}_{n+1})\beta_{n+1}(j)}{\sum_{\mathrm{n}=1}^{N-1}\sum_{j=1}^{L}\alpha_n(i)A_{ij}b_j(\mathrm{y}_{n+1})\beta_{n+1}(j)} \tag{3.29}$$

也就是：

$$A'_{ij}=\frac{\sum_{n=1}^{N-1}\alpha_n(i)A_{ij}b_j(\mathrm{y}_{n+1})\beta_{n+1}(j)}{\sum_{\mathrm{n}=1}^{N-1}\alpha_n(i)\beta_n(i)},\quad \begin{matrix} i=1,2,\cdots,L \\ j=1,2,\cdots,L \end{matrix} \tag{3.30}$$

此时，已经求出了 A'_{ij}。接着求a的估计。设a'为a的新的估计，因为 $n=1$ 时：

$$a'_i=\gamma_1(i) \tag{3.31}$$

也就是：

$$a'_i=\frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{L}\alpha_1(i)\beta_1(i)},\quad i=1,2,\cdots,L \tag{3.32}$$

得到了a的新的估计后，求B的估计，B表示系统的概率分布函数矢量，概率分布函数矢量B为一矩阵，b_{jm} 代表 HMM 系统在 S_j 时，观察到 $V_m, m=1,2,\cdots,M$ 的概率，这个 V_m 表示码字，M 是y的维度数，设 b_{jm} 的新的估计为 b'_{jm}。则：

$$b'_{jm}=\frac{\sum_{n=1:y_n=V_m}^{N}\gamma_n(j)}{\sum_{n=1}^{N}\gamma_n(j)} \tag{3.33}$$

从而 HMM 新的 $\lambda'=f(\mathrm{a}',\mathrm{A}',\mathrm{B}')$ 已经求出来了，然后当 $P(\mathrm{Y}\mid\lambda')>P(\mathrm{Y}\mid\lambda)$ 时，需要继续重复之前的步骤，再逐渐地精确 HMM 的这 3 个参数。

3.3 本章小结

本章通过介绍离散时域的有限状态自动机来陈述了 HMM 的基本原理，并引出了 HMM 在语音识别应用中需要解决的三个问题：概率计算、最优状态序列搜索以及参数估计。并对每一个问题的解决方法进行了详细讲解。

第 4 章 深度学习模型

研究表明，在机器学习需要表达高层次抽象（比如视觉、语音和其他人工智能任务）的复杂函数时应该使用深度学习算法。深度学习具有一个多层次非线性的操作，例如一个含有很多隐含层的神经网络，或者反复使用许多个子公式的复杂命题公式。求得深度学习的各层参数是一件具有挑战的工作。目前国内外提出了很多深度学习算法，例如递归神经网络(Recurrent Neural Networks, RNN)、深度信念网络(Deep Belief Network, DBN) 、限制玻尔兹曼机(Restricted Boltzmann Machines, RBM)等学习算法，使得解决这种问题更加容易，本章将对深度学习原理与方法进行介绍，其中将重点讨论在语音识别系统中取得显著成果的递归神经网络以及递归神经网络的改进结构--长短时记忆模型(Long-short Term Memory, LSTM)。

4.1 深度学习简介

让计算机学习人类生活的世界是近百年来人工智能科研人员的研究重点。为了达成这个目标，目前计算机已经接受了关于人类世界的很多信息。之前需要科研人员手工的将这些信息变换成计算机理解的形式，从而使得计算机用学习到的知识回答新环境下的问题。然而随着数据的增多，把所有的信息都手工标注是一件不可能完成的事情。因此，近年来许多科研人员纷纷研究学习算法，通过这些学习算法可以让计算机自己来获取人类输入信息中的有用信息。现阶段科研人员已经研究出了有很多成功的学习算法，并将这些算法应用到了人工智能的各个方面。但距离计算机真正地达到成年人的智力还是需要科研人员付出很长一段时间的努力的，以语音识别为例，目前世界上还没有一个能够非常流畅的和人进行沟通的学习算法，最好的自然语音算法也会在被人类问到一些涉及上下文相关、缩略主语等需要机器进行推理、思考等问题的时候出现理解错误。目前这种状况也出现在视觉、机器人系统以及其他人工智能级别的任务当中。所以推进现有的语音识别学习算法，使得那种语音识别学习算法能够像人类一样的思考、学习是当前人工智能的重要任务。

2006 年，从 Hinton 提出深度学习的算法以来，深度学习理论得到了大量的研究和应用。深度学习其实是机器学习的一个研究方向，是一种类似人脑神经系统中丰富层次结构的算法，它通过建立这种分层模型结构来对输入的数据进行逐层的提取，从而得到更高层次的抽象表示，这些抽象表示代表数据的特征和所属类型。深度学习算法通过多层非线性的信息处理来对数据进行特征提取、分类、模式分析和转换[39]。目前很多的深度学习算法通过无监督的学习形式出现，所以深度学习算法可以被应用在无标签的数据当中，这是其他算法无法比拟的，因为无标签的数据很容易获得，而且不用再人工进行标注，这项能力成为了深度学习的优势。

2011 年，Microsoft 在他们的语音识别系统引入了深度学习算法，从而使得这个系统的错误率降低了 20%-30%，这是 10 年来语音识别领域最大的一次改进。

2012 年产生了有很多深度学习方向的研究和应用，知名生物制药公司默克将深度学习算法使用到了分子药性预测的问题上来，并取得了世界上最好的效果。同年，Google 将深度学习算法带入到了 Google Brain 项目当中，这个项目构建了一个拥有 10 亿多个节点的自主学习神经网络。这个神经网络能够从大量的数据中归纳出事物的概念，这个网络后来被应用在无人汽车和图片搜索领域，并获得了巨大的成功。

2012 年 12 月，微软亚洲研究院展示了一套中英文即时翻译系统，这个系统的错误率只有 7%，而且翻译后的发音带有原发音人的音色，同时发音十分流畅。

2013 年，欧洲委员会开启了一项模拟人类大脑的超级计算机项目，该项目计划用时 10 年，投入 16 亿美元，全球 80 个相关机构共同参与，他们希望可以在研究人类大脑的工作方式上取得重要的进展，并且带动新兴计算机的研发。

目前，深度学习算法在国内外普及开来主要有以下几种原因：首先，计算机芯片的处理能力大大提高了，特别是计算的图像处理单元性能的提升；其次，计算机硬件的成本大幅度降低；其三，信号处理、机器学习、模式识别等研究的共同进步带动了深度学习算法的研究，因为这些研究的进步使得深度学习能够更好的利用集成、复杂的非线性函数来学习分布式的特征表示，并有效的利用无标签和有标签的数据。

4.2 深度学习和浅层学习

目前，国内外对机器学习的历史划分并没有达成一致，从不同角度来看，可以将机器学习的历史分为不同的阶段，如果按照机器学习模型的层次结构来划分的话，机器学习可以分为这两种：浅层学习(shallow learning)与深度学习(deep learning)[40]。

大部分信号处理技术和机器学习技术都可以被分为浅层结构[41]，比如最大熵(Maximum Entropy Models, MaxEnt)模型、支持向量机(Support Vector Machine, SVM)、Logistic 回归、核回归、条件随机场(Conditional Random Field, CRF)、高斯混合模型(Gaussian Mixture Model, GMM)、多层感知机(Muti-layer Perceptron, MLP)等都是浅层结构。这些结构一般只包含一层或者两层的非线性的特征变换，所以它们被看成是一个具有单隐含层或者根本没有隐含层的结构。这些浅层结构在处理比较简单的问题时很有效。但是这些算法只有有限的表征能力和建模能力，这使得它们不能够处理现实生活中绝大多数复杂的数据，比如自然图像、视觉场景、人的语音等数据。

而深度网络是一个包含多层隐含层的网络结构，科研人员在系统中引入了深度网络后，系统通过学习一种深层次的非线性网络来对一些复杂的函数进行逼近，从而这个系统能够提取出那些更复杂的特征[42]，所以深度网络有更好的表达能力。

深度网络也就是深度学习，所以深度学习相比传统浅层网络的主要优势就在于它能

够通过简单的方式来表示很多的函数集合。因为浅层结构难以表达复杂函数。如图 4.1 所示。

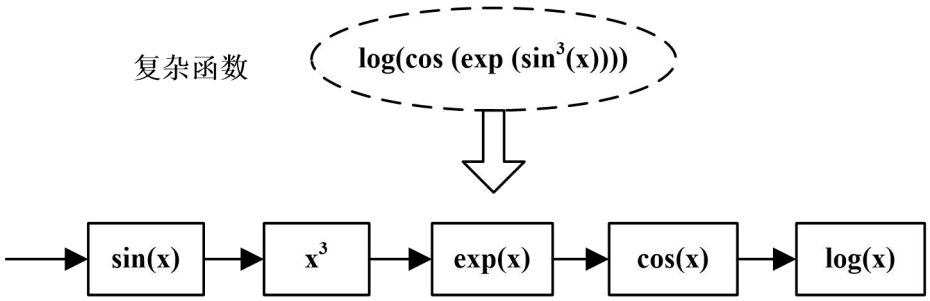

图 4.1 使用简单函数表达复杂函数的方法图

在训练深度网络时，需要将每层隐含层的激活函数设置为非线性，因为深度网络结合时就是将各层函数用线性函数组合起来。而隐含层的激活函数也是线性函数的话，这个深度网络就相当于一个单隐含层网络了，并没有增加网络的表达能力。

在科研人员处理一幅图像的时候，如果使用深度网络的话，一般会让网络学习“整体--部分”这种分解关系。也就是让计算机把一张大图片分成若干个小图片来看。

其实深度学习网络就是一个带有很多层隐含层的深度网络，这种深度学习网络一般使用大数据来进行训练。计算机可以得到很多有用的特征，从而可以提升计算机对数据分类和预测的准确性。所以说深度学习是特征学习里面一项未来前景很好，并且已经取得令人瞩目效果的一项技术。

4.3 深度学习的结构

深度学习网络是一种至少含有一层隐含层的网络，它能够为复杂的非线性系统建立模型，其多层次结构使得计算机具有更好的表达能力。目前深度学习有以下 3 种类型[43]：

（1）生成性深度结构

生成性深度结构一般用于计算观测数据和相关类型的联合概率分布、揭示数据的高阶相关特性。代表模型有深度信念网络、递归神经网络等。

（2）区分性深度结构

这种网络结构能够将模式进行分类，并计算数据的后验概率。这种结构代表模型是卷积神经网络(Convolution Neural Networks, CNN)，CNN 是现在众多科研领域的研究热点，这种网络在处理图像时，因为这种网络不用对输入的图片进行预处理，所以深受科研人员的喜爱。

（3）混合型结构

这种结构是第一种结构和第二种结构的混合。这种网络结构利用生成性结构作为输出，从而有益于网络结构的优化。

4.4 递归神经网络

递归神经网络(Recurrent Neural Network, RNN)属于生成性深度结构并拥有“记忆”的功能。因为语音信号的序列在时间上具有连续性，前一个基元和后一个基元有联系，并不是相对独立的，所以语音识别更适合使用 RNN 模型。但是一般的 RNN 所拥有的记忆时间并不是很长，会将当前时间点一段时间之前的信息忘记，也就是会出现“消失梯度”的现象。所以本文使用了一种改进的 RNN 模型--长短时记忆模型(Long Short Term Memory，LSTM)，这种模型加入了长短时记忆单元，这使得它具有了长时记忆的能力，从而能够更好的处理语音信号来进行语音识别。

4.4.1 多层感知器

半个多世纪前，Mcculloch 受到生物神经网络的启发，想要让计算机模拟人类学习、记忆、模仿的能力，首次构建了神经元模型，如图 4.2 所示。

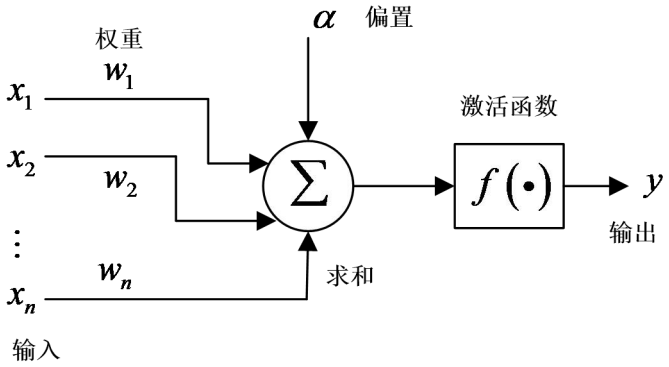

图 4.2 神经元模型

多层感知器就是一种单向多层的前馈神经网络，如何训练感知器的隐含层一直是一个难题，但是经过科研人员不断的探索，反向传播(Back Propagation，BP)算法有效的解决了这个问题[44]。递归神经网络通常含有一个多层的感知器，而且带有反馈回路，这将整个网络变成了一个映射器。

多层感知器(Multilayer Perceptron)是前馈神经网络(Feed-forward Neural Network)[45]，它的结构如图 4.3 所示。在输出层和输入层之间有一个以上的隐含层。这些步骤也就是前向传播(Forward Pass)。因为前向传播的输出和目标输出往往是有差距的，所以每次对比它们结果计算误差，并向反方向对权重进行修改，这样的过程被称为反向传递(Backward Pass)。

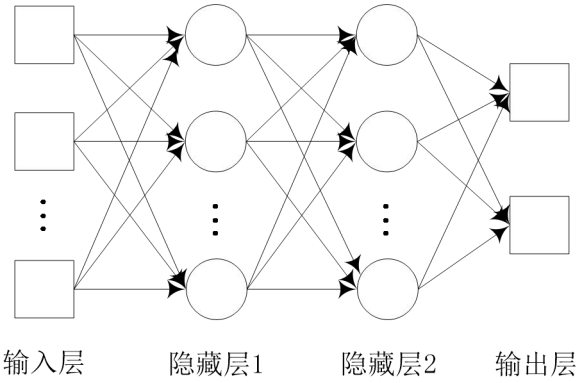

图 4.3 多层感知器

多层感知器输出结构只和当前的输入有关，和过去将来的输入信息无关。因此这些多层感知器适合对图像来进行分类，但是这种多层感知器并不善于对含有时间信息序列的数据进行处理。这就是本文为什么要将带有“记忆”功能的 RNN 引入到语音识别系统当中的意义。

4.4.2 递归神经网络

图 4.4 是一种递归神经网络(Recurrent Neural Network, RNN)[46]。这是一种用循环方式把前馈神经网络连接起来的网络。

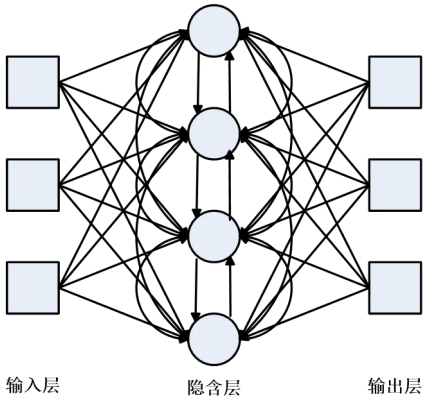

图 4.4 RNN 结构图

RNN 能把之前所有的输入都映射到每个输出的结果上，图 4.5 是 RNN 在时间方向上的简单展开结构。

如图 4.5 所示，每一个节点都表示这个时间点上的某一层神经网络的神经元。输入层与隐含层、隐含层到隐含层、隐含层与输出层的权重分别为“w_1”、“w_2”和“w_3”。这些权重会在每个时序上反复使用[47]。但是，一般的 RNN 会在每一次的反馈过程中丢失一部分信息，所以随着时间的积累，最原始的信息就会逐步消失，也就是常说的梯度消

失效应[48]，如图 4.6 所示。

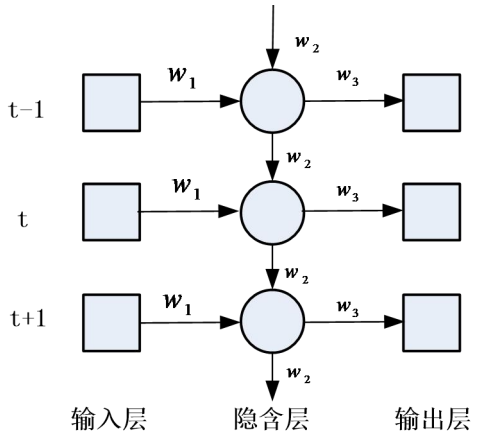

图 4.5 RNN 时间方向展开结构

图 4.6 中的每一个节点的颜色明暗程度都代表这个网络在这个时间点上对原始信息所记忆的程度。显而易见，伴随着新的信息的输入，隐含层关于原始信息的记忆被一次次的覆盖，最终“忘记”最原始的输入信息。

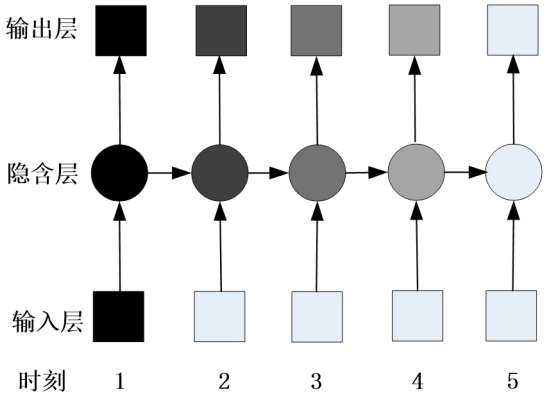

图 4.6 递归神经网络的梯度消失现象

4.4.3 长短时记忆网络

由上文得知，因为 RNN 具有消失梯度问题，所以很多学者做出了各种尝试，包括基于非梯度的训练[49]、时滞网络[50]、分层序列压缩[51]和回声状态网络[52]。最有效的技术之一就是由 Hochreiter 和 Schmidhuber 引入的[53]一种具有与 RNN 相同拓扑结构的长短时记忆模型(Long Short Term Memory, LSTM)。LSTM 网络都能够在一段较长时间的线性存储单元格中存储信息，从而克服消失梯度问题。

近年来，LSTM 技术已经成功的应用在很多模式识别任务当中，包括音素分类[54]、

情感识别[55]、手写识别[56]和驾驶员分心检测[57]。

LSTM 就是将 RNNs 的隐层神经元替代成记忆单元的网络。类似于 RNNs 中的循环连接，这些记忆单元循环连接。每个记忆单元由连接自身的记忆神经元和三个乘法门单元（输入、输出和遗忘门）组成。因为这些门单元允许在记忆单元内进行写入，查阅，和重置操作，一个 LSTM 单元可以看作是在数字计算机中的记忆芯片（但有不同）。门单元的整体效果是 LSTM 记忆神经元可以存储和访问一个跨越长时段的信息，从而避免消失梯度问题。例如，只要输入门保持关闭（相当于一个输入门激活接近于零），这个神经元的激活将不会被新的输入覆盖，因此序列打开输出门很久以后，这个神经元的激活才被用到。这允许在相关输入和输出间存在长时间的滞后，这可能不是标准的 RNNs，f_i，f_g，f_o 分别表示激活函数。图 4.7 表示一个包含了一个记忆神经元的记忆单元结构。

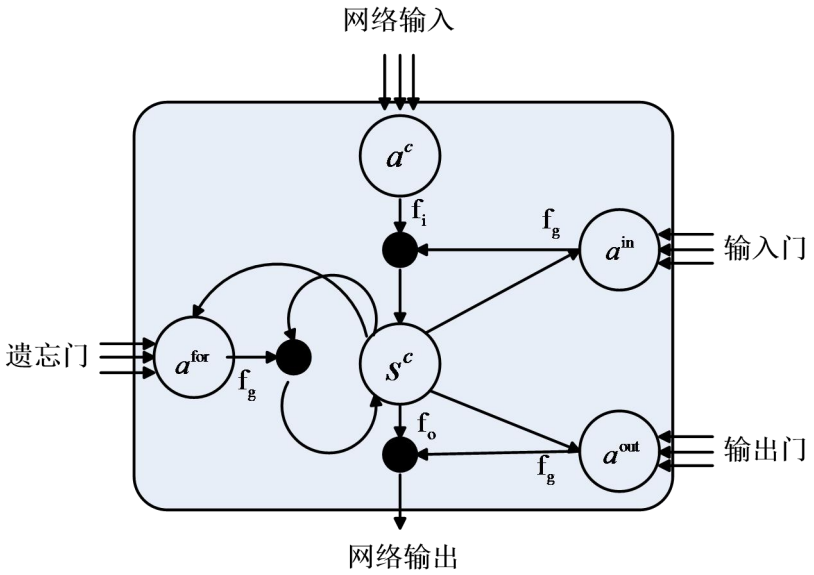

图 4.7 LSTM 的记忆单元结构图

如果 α_t^{in} 表示 f_g 被应用之前的时刻 t 的输入门激活。β_t^{in} 代表在应用激活函数后的激活，这个确定的记忆单元输入门激活（前向传播）分别被写为：

$$\alpha_t^{in} = \sum_{i=1}^{I} \eta^{i,in} x_t^i + \sum_{h=1}^{H} \eta^{h,in} \beta_{t-1}^h + \sum_{c=1}^{C} \eta^{c,in} s_{t-1}^c \tag{4.1}$$

和：

$$\beta_t^{in} = f_g(\alpha_t^{in}) \tag{4.2}$$

由于方程 4.2 是指 LSTM 网络中的一个特定的记忆单元，所有的变量都是标量。变量 η^{ij} 对应于从单元 i 到单元 j 的连接权重，此时，‘in’、‘for’和‘out’分别参照输入门、遗忘门、输出门。此时 H 代表隐含层的神经元数目，I 代表输入的数目，C 代表一个单元

的记忆神经元数目。s_t^c 代表 t 时刻神经元 c 的状态。

同样，在应用 f_g 之前和之后遗忘门的激活可以被计算，方法如下：

$$\alpha_t^{for} = \sum_{i=1}^{I} \eta^{i,for} x_t^i + \sum_{h=1}^{H} \eta^{h,for} \beta_{t-1}^h + \sum_{c=1}^{C} \eta^{c,for} s_{t-1}^c \tag{4.3}$$

$$\beta_t^{for} = f_g(\alpha_t^{for}). \tag{4.4}$$

这个记忆神经元的值 α_t^c 是一个在 t 时刻的输入和 t-1 时刻的隐含单元激活的加权和。

$$\alpha_t^c = \sum_{i=1}^{I} \eta^{i,c} x_t^i + \sum_{h=1}^{H} \eta^{h,c} \beta_{t-1}^h. \tag{4.5}$$

为了确定一个单元 c 的当前状态，首先需要测量遗忘门激活的先前状态和输入门的激活输入 $f_i(\alpha_t^c)$。

$$s_t^c = \beta_t^{for} + \beta_t^{in} f_i\left(\alpha_t^c\right). \tag{4.6}$$

输出门激活的计算方法与计算输入门、遗忘门激活时的方法一样，但是这次考虑当前状态 s_t^c，而不是先前时间步长的状态。

$$\alpha_t^{out} = \sum_{i=1}^{I} \eta^{i,out} x_t^i + \sum_{h=1}^{H} \eta^{h,out} \beta_{t-1}^h + \sum_{c=1}^{C} \eta^{c,out} s_t^c \tag{4.7}$$

$$\beta_t^{out} = f_g(\alpha_t^{out}) \tag{4.8}$$

最终，记忆单元的输出被定义为：

$$\beta_t^c = \beta_t^{out} f_o\left(s_t^c\right). \tag{4.9}$$

图 4.8 提供了在时刻 t-1 和时刻 t 的“展开”LSTM 网络连接的概况，为了简单起见，该网络只包含小的输入层与输出层，都只有两个节点，而且每个神经元只有一个记忆单元。LSTM 结构的最初版本只包含输入门和输出门。遗忘门是后来 Gers 所增加的[58]，从而允许记忆单元在网络需要忘记过去输入信息的时候重置自身。在本文后续实验中将使用包含遗忘门的增强 LSTM 版本。

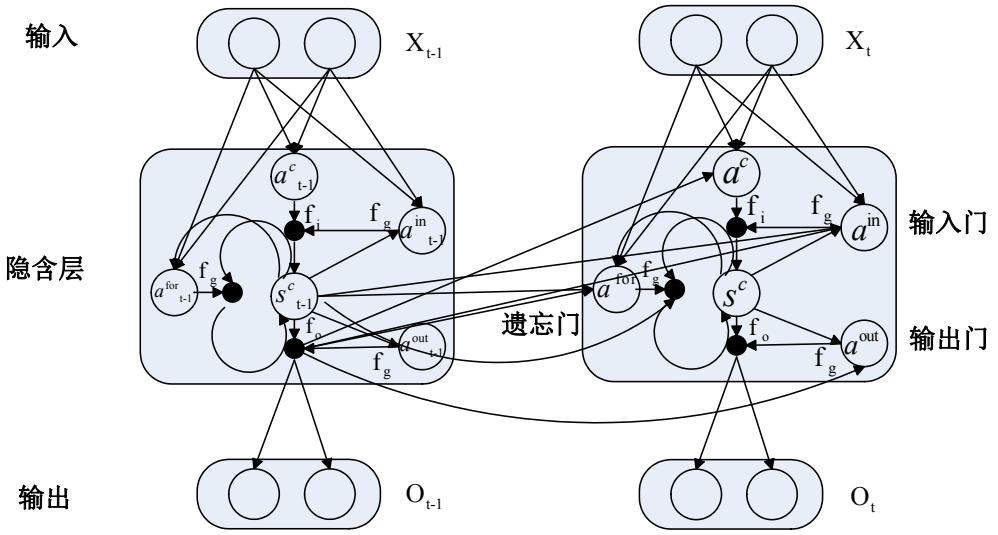

图 4.8 简化的 LSTM 网络连接

4.5 本章小结

本章首先对深度学习进行了简单介绍，对比了深度学习与浅层学习，突出了深度学习的优势。然后介绍了深度学习类型分类，分别为：生成性深度结构、区分性深度结构、混合型结构。其次通过介绍多层感知器的原理引出递归神经网络。在阐述了递归神经网络的优点与缺点后，提出了递归神经网络的消失梯度问题的解决办法----将 RNN 的隐层神经元用长短时记忆单元替代，网络变成长短时记忆网络，并对该网络进行了详尽介绍，长短时记忆网络将在本文中充当藏语声学特征提取器的角色。

第 5 章 LSTM-HMM 模型的藏语语音识别实验

本章首先对藏语发音和藏字进行了简单介绍，然后对本论文所使用的语料库建立方法进行了说明。其次介绍了本文的具体工作：在专业录音设备的帮助下对设计的语料库进行了录制、切分与标注。将这些藏语语音进行数字化处理、预处理后提取出 39 维 MFCC 特征向量。这些特征经过基于长短时记忆网络的藏语声学特征提取器提取后，生成 51 维的输出激活，然后将这 51 维的输出激活与 MFCC 特征结合后，应用 PCA 技术降维，提取最重要的 40 维 Tandem 特征输入给 HMM-GMM 模型，然后 HMM-GMM 模型再进行训练以及识别，最后对实验结果进行了总结。

5.1 藏语发音介绍

藏语是一种历史悠久的少数民族语言，使用藏语的人口众多，分布地广，而且藏文的建立伊始就有了大量的文献记载。所以藏语在我国的少数民族语言中具有举足轻重的地位，研究藏语将对促进民族间交流、保存中华文化起到一定的积极作用。

目前我国的西藏、云南、四川、甘肃等省共有 450 万人左右在使用着藏语。印度等中国周边邻国也有很多人在使用藏语。但是使用藏语的人口分布的比较广，所以藏语分出了很多种方言，主要有卫藏、康和安多这 3 种方言[60]。卫藏方言中的拉萨方言使用最为广泛，而且卫藏方言相当于汉语中的“普通话”，最为正统、官方。本文搭建的藏语语音识别平台的标准音定为卫藏方言中的拉萨方言。

5.1.1 藏文的介绍

从创建藏文开始便有很多的文献记载了藏族的知识、民间传说、文化、历史等，所以藏文记载和传承了很多藏族文化。藏文一般通过藏语拼音来对其进行拼写，并且书写藏语时，应该像汉语一样从左到右的顺序横向的书写，且藏文是一种二维结构的拼音文字，每个藏字表示一个藏语音节。

藏语拼音在音节中的位置有这两种，如图 5.1 所示。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。元音和后加字组成“韵母”其他字组成“声母”。图 5.2 是一个藏字的组成示例，这个藏字是藏语的“遮”。

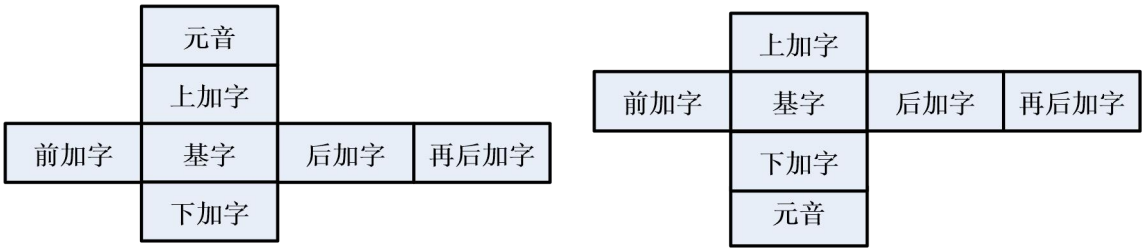

图 5.1 藏语音节的组成结构

图 5.2 藏字示例的组成结构

5.1.2 藏语拉萨方言拼音的声韵母

藏语拉萨方言一般由西藏拉萨市附近的藏民使用，而且藏语拉萨方言也被定义为藏语的标准发音，拉萨方言可以说是藏族人民的“普通话”。本文将藏语拉萨方言作为研究重点。藏语拉萨话在发音上有很多独特的地方，比如，声母中没有浊音以及阻塞音；并且复辅音声母也不常见；具有真性复合以及鼻化元音；声调也没有像汉语声调一样有很多的起伏，发音很平稳；

藏语拉萨方言与传统的藏语相比，在发音上做出了很多的改变，它将传统藏语中的复辅音化简成为单辅音，而且将传统的浊音阻塞音做了清化的改变，所以说，藏语拉萨方言里没有浊音阻塞音声母以及复辅音声母。另外，藏语拉萨方言还产生了一种新的辅音声母，它的韵母也较传统藏语而言做出了很多的改变，因为传统藏语有五个元音，但是藏语拉萨方言却有八个单元音以及几个鼻化元音。目前拉萨方言中保留着传统藏语的流音韵尾、鼻音、塞音。拉萨方言的口语与书面用语都不相同，在口语中，清辅音是与送不送气相对立的而不是与清浊相对应的，这一点和汉语普通话很像。不仅如此，同汉语普通话一样，拉萨方言的口语中还有 4 个卷舌音，分别是 / r / 、/ sh / 、/ ch / 、/ zh /

在发音上，拉萨方言的声母含有复辅音以及单辅音，一般常用单辅音声母。藏文的辅音能够独自组成音节，它的发音就被看成它的拼音名称。而当它的辅音独立成音节时，它的发音全有/a/和固定声调。辅音中 16 个为低调，14 个为高调。它的声母中有 28 个辅音音节，根据发音部分以及发音方法，可将这 28 个辅音分为如下几类，见表 5.1。

拉萨方言中的韵母含有 8 个元音，以及/r/、/g/、/b/、/m/的组合音。根据拉萨方言的发音方法和发音部分，可以将拉萨方言的韵母分为几类，如表 5.2 所示。

表 5.1 根据发音部分以及发音方法，藏语拉萨方言的声母分类表

	双唇	唇齿	舌尖前	舌尖中	舌尖后	舌面前	舌面中	舌根	喉
塞音不送气	b			d			gy	g	
塞音送气	p			t			ky	k	
塞擦音不送气			z		zh	j			
塞擦音送气			c		ch	q			
擦音		F	s	lh	sh	x	hy	h	h
通音					r				
鼻音	m			n		ny	ng		
边音				l					
浊音	w						y		

表 5.2 根据发音方法以及发音部位，藏语拉萨方言的韵母分类表

	开口呼	齐齿呼	合口呼	噘口呼
单元音	a,o,e,aa,ee,oo,ae,ue,oe	i,ii ,ie,	u,uu ,uo	yu,yue
双元音	Au	iou(iu)		
鼻音	aen,en,oen,on,ang,eng	in,ing	Un	
后接 m	am,em,om	Im	Um	
后接 b	ab,eb,ob	Ib	Ub	
后接 g	ag,eg,og	Ig	Ug	
后接 r	ar,er,or	Ir	Ur	

5.1.3 藏语的声调

藏族人民一般用藏语来进行沟通与交流，所以与汉语普通话一样，藏语的声调同样也很重要，一般可以根据藏语的语调来判别语音与词意。不一样的藏语方言有不一样的声调个数，安多方言没有声调，木雅方言有 2 个声调，工布方言有 6 个声调，本文的研究对象藏语拉萨方言有 4 个声调。

拉萨方言的这 4 个声调具有高低的分别，当它的韵尾类型是舒声韵时，音调变高，当韵味类型是粗声韵时，声调边低。而且，拉萨方言中的元音长短也能对声调产生作用。拉萨方言的调值分别为 53、55、12、14。

5.2 语音样本库的建立

目前，藏族地区的电视、广播电台等媒介已经将拉萨方言定义为藏语的标准音来进行播报，接下来本文将对这种语言进行韵律建模，设计这种语言的标注格式。

语音文件的标注格式一般有：（1）语料的选择；（2）语音的录制；(3）文本和语音的言语标注；（4）数据的管理。

5.2.1 文本语料库的设计

想要评测一个语音识别系统的好坏就需要将其应用到实际当中来，接受真实数据的检验。语音识别系统所需要的数据就是语音信号，用什么样的语音数据对语音识别系统进行检测，这是十分关键的一步，所以应该合理的设计语料库。

应该用合理的方法来选取藏语文本语料。首先这些字应该能够在在发音上体现藏语拉萨方言的声母、韵母、音调上的特性。然后选取的这些藏语语料应该非常常见，这样训练好的语音识别系统能够有更广的使用范围。

本文通过对藏语报刊、杂志以及藏语的朗读文章进行查阅与总结后，选出了 51 个比较常见的藏语文字，这 51 个藏语文字都是藏语拉萨方言中的单音节字，这些语料包括藏民经常接触的藏文数字（一到十）、一些常见藏文名词（21 个）、动词（11 个）、其他词性藏字（9 个），构建的文本语料如表 5.3 所示。

5.2.2 语音语料的录制

录制语料时，本文聘请了 4 个人对这 51 个藏字进行朗读，其中 1 人为女生，3 人为男生，年龄分别在 18-25 周岁，发音标准，清晰。这 4 人对这 51 个藏字每个读 30 遍，共 6120 个样本。每个人录制的语料中，3 个发音为一组，一组中的第一遍和第三遍作为训练语料，总共 4080 个样本。第二遍作为测试语料，总共 2040 个样本。在录音的时候为了保证录音的质量，使用专业的高保真录音话筒以及外置声卡在专业的高隔音录音棚内录制，录音电脑为四核版的 Mac Pro。在录音的过程中，录音棚内采用计算机屏幕提示系统，这样很大程度上减少了录音过程中的可能发生的噪声。本文将录音语句以 16kHz 的采样率和 16 位采样精度的形式保存为单声道的 Microsoft WAV 格式。图 5.3 是本文使用的录音软件 PRO tools 的工作界面。

在每一次录音结束后，我们会用专业的监听耳机对录制好的语音信号进行检测，以防有的语音漏录或者说话人说错，多说。如果出现相关问题，我们将对问题语音进行补录以及删除。

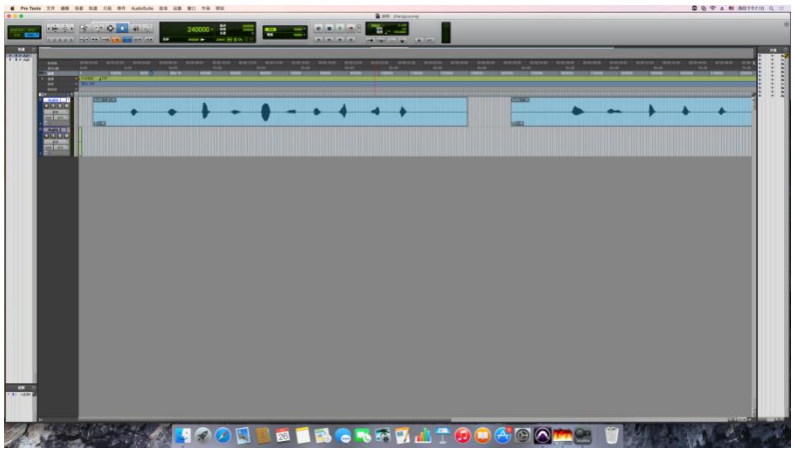

图 5.3 录音软件 PRO tools 的工作界面

表 5.3 文本语料中的所有藏字

藏字	翻译	藏字	翻译	藏字	翻译	藏字	翻译	藏字	翻译
གཅིག	一	དྲུག	六	ཞབས	脚	ཀང་ཟྭབས	袜子	འཁྱགས	冷
གཉིས	二	བདུན	七	ཚིགས	关节	ལེངས	充满	གངས	雪
གསུམ	三	བཀྲད	八	སྐྱ ག	读	གོང	他	གོང	以上
བཞི	四	དགུ	九	བཀོག	摘	འཁོན	挂上	ཤར	东
ལྔ	五	བཅུ	十	བཀྱག	举	འཁོང	转动	ཡའི	也
འཇིབས	阴沉	ཨ་ཡ	叫	ཀ	茶	སྐྱེབས	到达	སྐད་ས	舌头
ཆད	断的	དྲངས	香	ཁྱི	狗	དབུ	头	ཚེམས	牙齿
ང	我	སྨན	药	བཟློག	弄丢	ལྟོག	眼睛	རླངས	肿大
ངའི	我的	ཆུ	水	གློག	电	ཤན་ས	鼻子	སྐྱོན	缺点
མཚན	名	ཆང	青稞酒	རྟ	马	ཁ་ཟ	口	བཀྲལ	煮
འག	下面								

5.2.3 语料的切分和标注

语料的标注对语料库很重要，一个标注准确的语料库使得科研人员在进行语音合成或语音识别时事半功倍。藏语中，藏字是以音节为单位的，一个音节就是一个单位，本文选择了 51 个单音节字作为本文的研究语料，所以切分出的每个音节都是一个藏字。本文使用音频处理软件 Cool Edit Pro 作为语料切分工具。图 5.4 是本文使用的 Cool Edit Pro 切分藏字录音 “ཨཀ”时的工作界面。

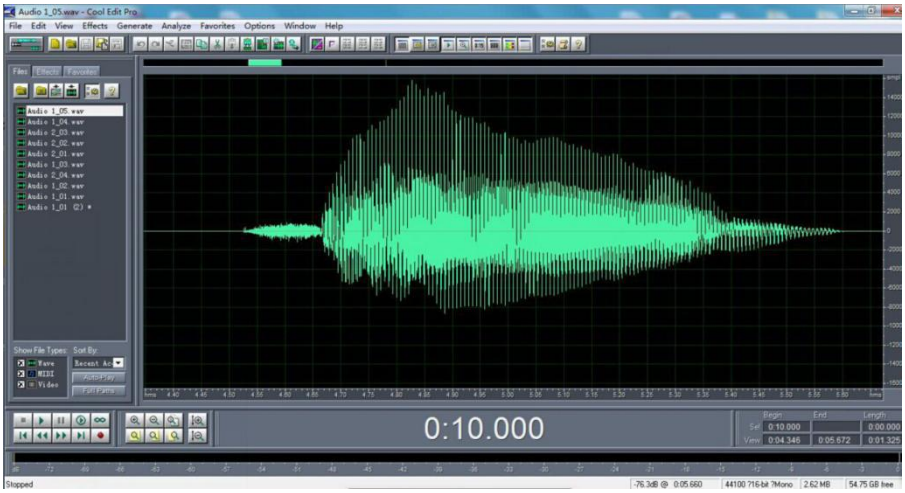

图 5.4 使用的 Cool Edit Pro 切分藏字录音“ཨཀ”时的工作界面

本文把语料库中的每个音节的起始点和结束点都标注出来。因为在这个语料库的文本语料设计上，本文选择了 51 个单音节字作为本文的研究语料，所以在标注每个音节的时候就相当于对每个藏字进行标注。本文语料库标注的主要工作包括音段和韵律的标注，所谓的音段标注是对拉萨方言中的声韵母进行标注，并标注出音素的定位，而有关音调的标注则是韵律标注。标注时，本文为每个待标注音素设计一个固定且好区分的字母或者组合来进行标注。在标注好每个藏字的音节信息后，本文需要对这些拉萨方言进行转写。转写时采用机读音标标注方案完成工作，然后根据藏语在语音发音上的特征，采用一套专门针对藏语语料库的标注方案对本文设计的语料库进行标注。

（1）标注方案设计

上个世纪 80 年代后期，欧洲信息技术研究中心设计一种可以直接使用键盘对语音的音标进行标注的标注方案，这个方案叫做机读音标(Speech Assessment Methods Phonetic Alphabet, SAMPA)。然后上世纪 90 年代中期，有研究人员将这个音标系统进行了扩展，通过这个扩展过的方案可以让相关的研究人员直接用 ASCII 码来对所有语音的音标进行标注。现在这种标注方案在全世界范围内通用，包括中国台湾的“国语”、香港和广东等地的“粤语”以及大陆使用的“普通话”等都有各自的 SAMPA 标注方案。

因为藏语和汉语普通话都是汉藏语系，都是由声母以及韵母所组成的，且发音相似，

所以它们的标注方案相似。目前普通话的标注方案已经被完整的设计出来，本文借助由普通话标注方案 SAMPA-SC，设计的藏语拉萨方言的标注方案 SAMPA-T 来进行语料标注，这里的 T 指 Tibetan 即藏语的意思。设计标注方案 SAMPA-T 的流程图如图 5.5 所示。

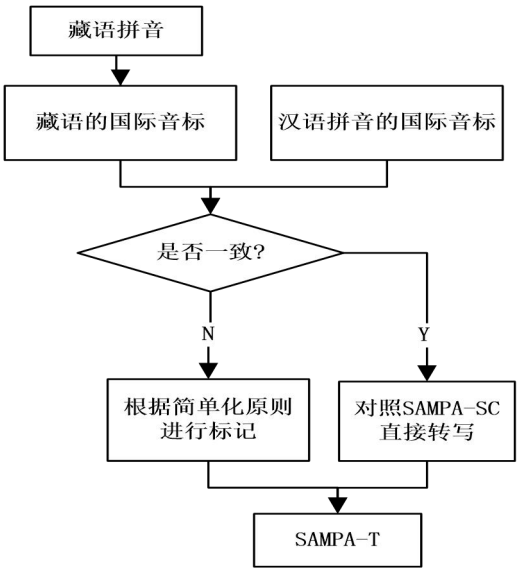

图 5.5 基于汉语机读音标 SAMPA-SC 的藏语 SAMPA-T 转写流程图

这种 SAMPA-T 转写的具体步骤就是依照这两种语言的国际音标进行转写。如果汉语拼音的国际音标中有藏语拉萨方言的国际音标，就直接使用 SAMPA-SC 标注拉萨方言。如果 SAMPA-SC 中没有藏语的标注方案，本文就使用自定义的符号来对藏语拼音进行标注，这些自定义的符号都是可以用 ASCII 码进行表示的。SAMPA-T 所自定义的符号与汉语普通话的标注方案 SAMPA-SC 所定义的符号不发生冲突，且在具有相同发音时，使用相同的标注方案，所以本文设计的这种标注方案可以在未来扩充到汉藏语同时识别的时候继续使用。

（2）藏语的 SAMPA-T 标注方案

藏语拉萨方言有 36 个声母、45 个韵母，SAMPA-SC 标注了 22 个声母、39 个韵母。两种语言有 20 个相同声母音标以及 13 个相同的韵母音标。所以它们在具有相同音标的声母、韵母标注上使用汉语普通话标注方案 SAMPA-SC 进行标注。而对于发音不同的声韵母，则采用自己定义的符号来对藏语拉萨方言进行标注。汉语声母与藏语声母具有相同国际音标时，标注方案如表 5.3 另外，汉语普通话的韵母和几个藏语拉萨方言的韵母也相同，所以这些藏语韵母也和普通话的 SAMPA 标注方法相同，如表 5.4。

表 5.3 汉语声母与藏语声母具有相同国际音标时的 SAMPA 对照表

藏语拼音	汉语拼音	国际音标	SAMPA-SC	SAMPA-T
G	G	[k]	G	g
K	K	[k^h]	K	k
J	J	[tɕ]	dz`	dz`
Q	Q	[$tɕ^h$]	ts`	ts`
D	D	[t]	D	d
T	T	[t^h]	T	t
N	N	[n]	N	n
B	B	[p]	B	b
P	P	[p^h]	P	p
M	M	[m]	M	m
Z	Z	[ts]	Dz	dz
C	Ch	[ts^h]	C	C
W	W	[w]	W	W
X	X	[ɕ]	s`	s`
S	S	[s]	S	S
Y	Y	[j]	Y	Y
L	L	[l]	L	L
Zh	Zh	[tʂ]	Zh	zh
Ch	Ch	[$tʂ^h$]	Ch	ch
Sh	Sh	[ʂ]	Sh	sh

但是汉语普通话的声母和藏语拉萨方言的声母有些是不同的，所以本文需要对藏语拉萨方言的音标进行自定义标注，表 5.5 是本文所自定义的一部分标注。

（3）语料的标注

因为语料的标注准确性对一个语音识别系统的训练效果有着至关重要的作用，所以需要很认真的来对待语料的标注。目前本文的实验语料切分都是由本人及同一研究课题的同学手工完成的。在切分工作完成之后，本文需要对切分好的藏语语料进行手工标注且仔细检查，这样可以保证语料的准确性。

表 5.4 汉语韵母与藏语韵母一样时的 SAMPA 对照表

藏语拼音	汉语拼音	国际音标	SAMPA 定义
A	a	[a]	A
O	o	[o]	O
I	i	[i]	I
U	u	[u]	U
Ü	ü	[y]	Y
Au	ao	[ɑu]	Au
Ang	ang	[ɑŋ]	An
Eng	eng	[əŋ]	En
Ung	ong	[uŋ]	On
Ing	ing	[iŋ]	In
Ie	ie	[iɛ]	Ie

表 5.5 汉语声母与藏语声母不一致时的 SAMPA 对照表

藏语拼音	国际音标	SAMPA 自定义	汉语拼音	国际音标	SAMPA 自定义
Ng	[ŋ]	ng	f	[f]	F
Ny	[ɲ]	ny	h	[x]	X
R	[ɹ]	r	v	[v]	V
H	[h]	h			
Gy	[c]	K1			
Ky	[c^h]	kh			
Lh	[ɬ]	lh			

5.3 语音数据特征提取

本实验中，预处理使用汉明窗、窗长 25ms、帧移 10ms。然后提取一组 39 维 MFCC 特征，这些特征包括 1 维的能量特征、12 维 Mel 倒谱系数特征、以及它们的一阶与二阶差分。实验中 MFCC 特征提取过程如图 5.6 所示。本文将这些 MFCC 特征作为 LSTM 的输入，从而输出 Tandem 特征。

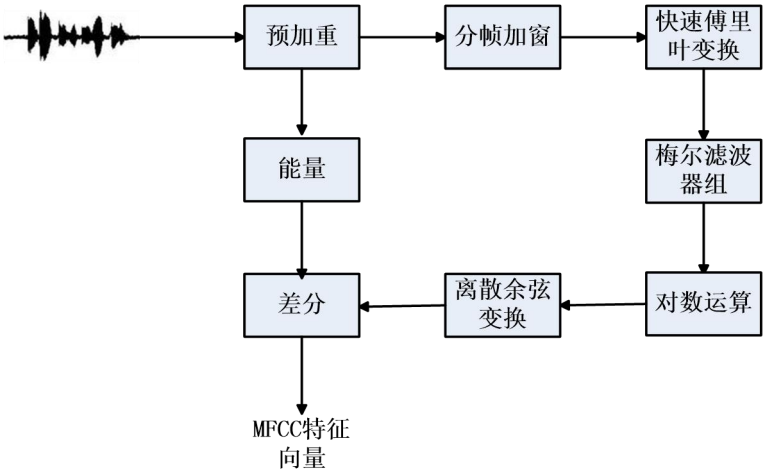

图 5.6 39 维的 MFCC 特征提取过程

5.4 递归神经网络配置

本实验使用 LSTM 网络来对 39 维的 MFCC 特征进行更近一步的处理。本文首先对实验所需网络结构进行配置，经过查阅文献以及实验，本文最后确定本实验所使用的网络结构配置如表 5.7 所示，本实验所构建的长短时记忆网络结构如图 5.7。表中第 1 列是该网络各层的名称；第 2 列代表该网络各层神经元的数目；第 3 行代表的是该网络各层神经元的类型；第 4 行是该网络各层的偏置数。

该网络输入的是特征提取阶段提取的 39 维 MFCC，生成的 51 维输出激活代表语料库中每个基元的后验概率，也就时 51 个字的后验概率。本文将这 51 维输出激活与 39 维的 MFCC 特征结合到一起生成 90 维特征。然后使用主成分分析(Principal Component Analysis, PCA)算法进行降维，提取最重要的 40 维 Tandem 特征输入给 HMM-GMM 模型，然后进行训练和识别。

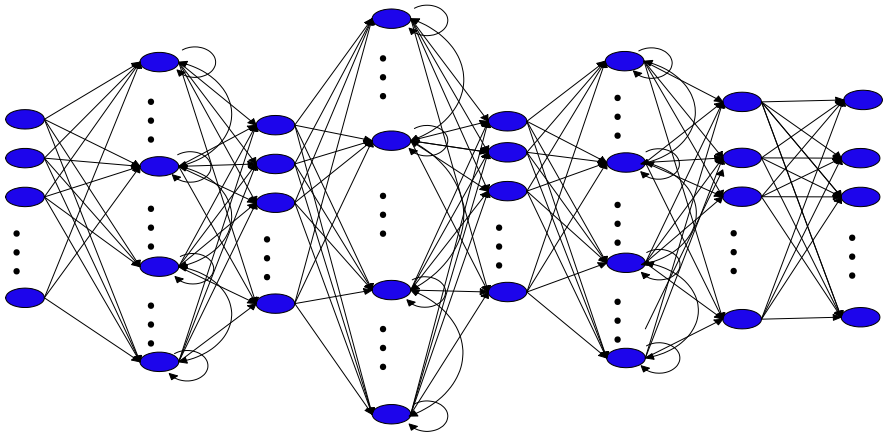

图 5.7 长短时记忆网络结构

表 5.7 LSTM 网络配置图

Name	Size	Type	Bias
input	39	input	
blstm_level_0	156	blstm	1.0
subsample_level_0	39	feedforward_tanh	0.0
blstm_level_1	300	blstm	1.0
subsample_level_1	75	feedforward_tanh	0.0
blstm_level_2	102	blstm	1.0
output	51	softmax	1.0
postoutput	51	multiclass_classification	

PCA 算法是一种对高维数据进行降维和去除噪声的方法。PCA 技术能使用较少的 M 维特征把原有的 N 维特征取代(M<N)。这种算法的具体处理过程如下：

第一步：特征中心化。每一维原始数据减去该维度上的均值得到的矩阵 Y。

第二步：计算矩阵 X 的协方差矩阵 Z。

第三步：计算矩阵 Z 的特征向量和特征值

第四步：将 Z 的特征值从大到小排序，需要多少维特征就保留前多少维特征向量。

本文将 90 维 Tandem 特征经过 PCA 技术降维后取最重要的 40 维特征向量输入给 HMM-GMM 模型进行训练以及识别。基于 LSTM 网络提取 Tandem 特征的流程图如图 5.8 所示。

在训练阶段，先将输入的藏语语音信号进行语音信号处理，提取出这段语音信号的 39 维 MFCC 特征，然后使用这些特征作为 LSTM 的输入，这个深度学习神经网络的输出是本文语料库中 51 个字的后验概率。将这 51 维输出激活与 39 维 MFCC 特征结合到一起生成 90 维特征。然后经过主成分分析算法对特征进行降维，提取前面 40 个 Tandem 特征输入给 HMM-GMM 模型进行训练。

在识别阶段，开始的实验步骤都相同，同样将藏语语音信号进行语音信号处理提取出 39 维 MFCC 特征，将这 39 维 MFCC 特征通过 LSTM 网络，将该网络输出的 51 维输出激活与 39 维 MFCC 特征组合生成 90 维特征，利用 PCA 技术提取最重要的 40 维特征。对照 HMM 模型库，输出最有可能的输出结果完成识别。本论文的流程图如图 5.9 所示

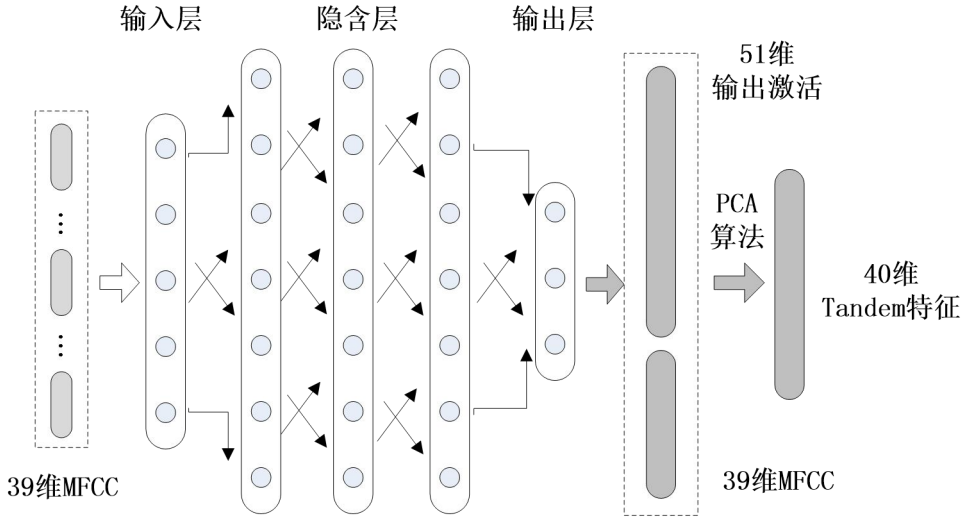

图 5.8 基于 LSTM 网络提取特征的流程图

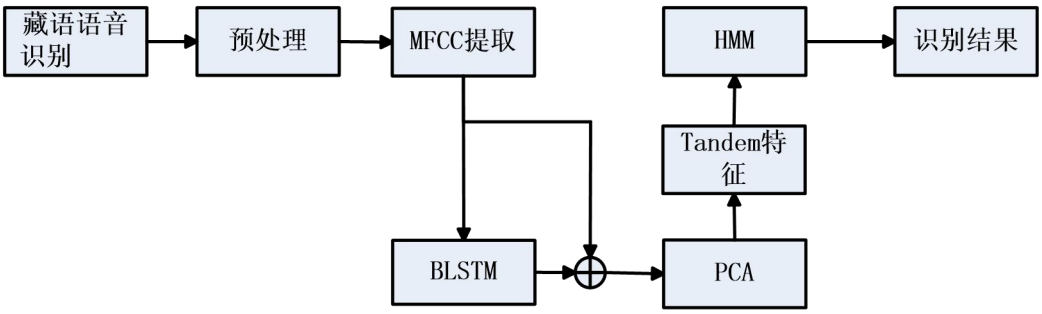

图 5.9 本藏语语音识别平台的流程图

5.5 实验结果

本实验在 Ubuntu14.04 64 位版本上完成，本文提出的这种结合深度学习模型与 HMM 的藏语语音识别平台的训练错误和测试错误如图 5.10 和图 5.11 所示。由图可以看出本藏语语音识别平台在到达 200 次迭代周期后，本实验的错误率趋于平稳。本实验训练错误曲线如图 5.10 所示，由图可以看出本平台在经过了 200 个迭代周期后，本实验的测试错误率趋于平稳，能够稳定到 0.96%以下。本实验的测试错误曲线如图 5.11 所示，由图可以看出本实验在经过了 200 个迭代周期后，测试错误率趋于平稳，能够稳定到 19.44%以下，也就是本语音识别平台能够在本文建立的面向藏语语音识别的藏语语料库的训练集中识别率稳定达到 99.04%以上，本语音识别平台能够在本文建立的面向藏语语音识别的藏语语料库的测试集中识别率稳定达到 80.56%以上。本文的结果证明基于长短时记忆网络的藏语声学特征提取器能够在藏语语音识别中起到积极地作用，本文工作将对促进民族间交流起到积极作用并为以后的藏语语音识别研究提供基础和借鉴。

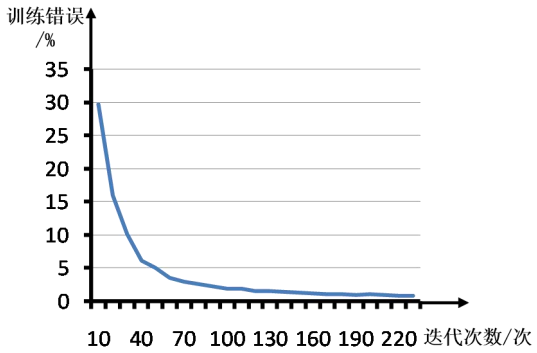

图 5.10 系统的训练错误曲线

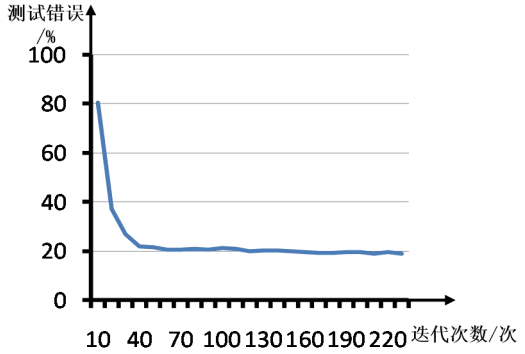

图 5.11 系统的测试错误曲线

5.6 本章小结

本章首先对藏语进行了简单介绍，其中包括藏字音节的组成结构、声韵母发音、声调等内容。然后介绍了本文语料库建立的具体办法，包括语料库的设计、录制与标注。随后通过介绍汉语机读音标的标注方案 SAMPA-SC 的标注方法，提出了藏语的 SAMPA-T 标注方案。接着介绍了本文实验特征参数的提取方法以及长短时记忆网络作为藏语声学特征提取器提取特征的方法。最后对本文实验结果进行了介绍，本实验在构建的藏语语料库的测试集中能够达到 80.56%以上的识别率，在藏语语料的训练集中能够达到 99.04%以上的识别率。本文工作将对促进民族间交流起到积极作用并为以后的藏语语音识别研究提供基础和借鉴。

第 6 章 总结与展望

6.1 论文总结

让机器听懂人类的话，并根据人类的命令完成工作，这是很多科研人员多年来的努力方向。近些年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，深度学习在各个方面得到了广泛的应用，并且已经在语音识别领域取得了惊人的成绩。所以近些年许多国内外研究机构都争先恐后的将深度学习算法引入到了他们的语音识别系统当中，其中 LSTM 网络是一种被证实了具有很好语音识别效果的深度学习结构。藏族是中国不可或缺的一个民族。我国现有 450 万人在使用着藏语。但是目前国内外还没有一个研究机构将 LSTM 网络引入到藏语语音识别系统当中。本文将 LSTM 模型与传统的 HMM 模型相结合，使用深度学习模型作为声学特征提取器为 GMM-HMM 模型提供更好的声学特征，从而提高识别率。

本文的主要研究内容与工作如下：

1.本文对藏语语音识别的研究现状和历史进行了总结。然后重点介绍了藏语语音识别系统的主要结构和藏语语音信号处理的具体过程。主要包括语音信号的数字化处理、预处理以及目前主流的几种特征提取方法。

2.重点研究了 HMM 模型与深度学习模型。包括 HMM 的原理、训练和识别的算法与具体过程。然后对深度学习算法的研究现状和基本原理进行了研究 。

3.在对比了藏语与汉语发音特点后，借助现有的汉语普通话标注方案 SAMPA-SC，设计了藏语拉萨方言的标注方案 SAMPA-T。并设计了一个包含 51 个常见藏字的文本语料库，并聘请了 4 名在校大学生进行藏语语料的录制，然后对实验语音进行了标注。

4. 深入研究了递归神经网络及其改进结构--长短时记忆模型的原理和算法，使用长短时记忆网络作为藏语语音识别系统中的声学特征提取器。利用 LSTM 网络产生出语料库中 51 个字的后验概率，并将这 51 维输出激活与 39 维的 MFCC 特征加到一起，然后经过 PCA 算法降维，提取前 40 个 Tandem 特征，然后将这些特征输入给 HMM-GMM 模型进行训练与识别。

5.实现了一个基于深度学习算法的声学特征提取器，利用深度学习模型与 GMM-HMM 模型各自的特点，将这两种模型结合了起来，应用在了藏语拉萨方言的语音识别系统当中。本项探索性的研究将会对促进民族间交流起到积极作用并为以后的藏语语音识别研究提供基础以及借鉴。

6.2 下一步的工作展望

经过了硕士阶段的学习与研究，搭建了一套基于深度学习网络的藏语声学体征提取器，并与 GMM-HMM 模型结合，实现了一套藏语孤立词识别系统，但是目前本系统还

有很多地方需要完善。为了取得更好的识别结果，接下来的工作将会在以下的几个方面进行展开：

1.语料库的建设

本系统的语料库是根据藏语常用词进行设计的，只包含 51 个孤立词。所以本语料库只适用于小语料的藏语语音识别系统。所以未来工作需要扩大语料库，并对连续语音、噪声语音进行设计与录制。

2.克服环境噪声

本论文所使用的语料的都是在专业录音棚内获得的纯净藏语语音，但是现实生活中的语音大多伴有各种噪声，所以本文中的藏语语音识别系统在实际应用中还不能取得最好的效果。所以研究相关的语音降噪算法，并应用到本文的语音识别中，能够增加噪声鲁棒性，使得该系统可以更好的在现实生活中使用。

3.bottleneck 特征

ANN 可以提取两种特征。一种是本文所使用的 Tandem 特征，一种是 bottleneck 特征。目前科研人员使用 Tandem 特征的比较多，因为 Tandem 特征与 ANN 的输出层关系紧密，所以输出层只能含有小量的单元数目。而 bottleneck 特征是将神经网络中的某一隐含层变为 bottleneck 层，然后计算神经网络的线性输出而提取的。大量实验证明，这种 bottleneck 特征可以提高大词汇量的语音识别系统的识别率，所以希望在以后的工作中，将 bottleneck 特征作为研究重点，并应用到本平台当中。

随着语音识别技术理论的不断创新以及计算机计算能力的不断提高，语音识别系统的识别效果将会越来越好，并应用到更多的领域当中，从而使得人们生活和工作更加便利，将各种新的语音识别技术引入到藏语的语音识别系统当中将会对促进民族间交流起到积极的作用。希望本文的工作可以为更多旨在促进藏语信息化发展的科研人员提供研究基础与借鉴。

参考文献

[1] Juang B H, Rabiner L R. Hidden Markov models for speech recognition[J]. Technometrics, 1991, 33(3): 251-272.

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

[3] Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning[J]. APSIPA Transactions on Signal and Information Processing, 2014, 3: e2.

[4] Wöllmer M, Weninger F, Geiger J R, et al. Noise robust ASR in reverberated multisource environments applying convolutive NMF and Long Short-Term Memory[J]. Computer Speech and Language, 2013, 27(3): 780-797.

[5] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]// Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 6645-6649.

[6] Fan Y, Qian Y, Xie F L, et al. TTS synthesis with bidirectional LSTM based recurrent neural networks[C]// Proc. Interspeech. 2014: 1964-1968.

[7] Davis K H, Biddulph R, Balashek S. Automatic Recognition of Spoken Digits[J]. Journal of the Acoustical Society of America, 1952, 24(6):669.

[8] Fry D B. Theoretical aspects of mechanical speech recognition[J]. British Institution of Radio Engineers, 1959, 19(4): 211-218.

[9] Denes P. The design and operation of the mechanical speech recognizer at University College London[J]. British Institution of Radio Engineers, 1959, 19(4): 219-229.

[10] Reddy D R. Approach to computer speech recognition by direct analysis of the speech wave[J]. Journal of the Acoustical Society of America, 1966, 40(5): 1273-1273.

[11] Vintsyuk T K. Speech discrimination by dynamic programming[J]. Cybernetics and Systems Analysis, 1968, 4(1): 52-57.

[12] Itakura F, Saito S. Statistical method for estimation of speech spectral density and formant frequencies[J]. Electronics and Communications, Japan, 1970, 53(1): 36.

[13] Jelinek F, Mercer R L, Bahl L R. 25 Continuous speech recognition: Statistical methods[J]. Handbook of Statistics, 1982, 2: 549-573.

[14] Schwartz R, Chow Y L, Kimball O, et al. Context-dependent modeling for acoustic-phonetic recognition of continuous speech[C]// Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1985: 1205-1208.

[15] Schwartz R, Chow Y L, Kimball O, et al. Context-dependent modeling for acoustic-phonetic recognition of continuous speech[C]// Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1985, 10: 1205-1208.

[16] Juang B H, Rabiner L R. Hidden Markov models for speech recognition[J]. Technometrics, 1991, 33(3): 251-272.

[17] Furui S. History and Development of Speech Recognition[M]. Speech Technology. Springer US, 2010:1-18.

[18] Rumelhart, David E, McClelland, James L, PDP Research Group, CORPORATE. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations[J]. Language, 1986, 63(4).

[19] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models[J]. Computer Speech and Language, 1995, 9(2): 171-185.

[20] Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition[J]. Computer Speech and Language, 1998, 12(2): 75-98.

[21] Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains[J]. IEEE Transactions on Speech and audio processing, 1994, 2(2): 291-298.

[22] Bahl L R, Brown P F, De Souza P V, et al. Maximum mutual information estimation of hidden Markov model parameters for speech recognition[C]// Proc. ICASSP, 1986, 86: 49-52.

[23] Brown P F. The acoustic-modeling problem in automatic speech recognition[M]. UMI, 1987.

[24] Valtchev V. Discriminative methods in HMM-based speech recognition[D]. University of Cambridge, 1995.

[25] Valtchev V, Odell J J, Woodland P C, et al. Lattice-based discriminative training for large vocabulary speech recognition[C]// Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996, 2: 605-608.

[26] Cardin R, Normandin Y, Mori R D. High performance connected digit recognition using maximum mutual information estimation[J]. IEEE Transactions on Speech and Audio, 1994, 2(2):299-311.

[27] Valtchev V, Odell J J, Woodland P C, et al. MMIE training of large vocabulary recognition systems[J]. Speech Communication, 1997, 22(4): 303-314.

[28] Juang B H, Katagiri S. Discriminative learning for minimum error classification[J]. IEEE Transactions on Signal Processing, 1992, 40(12): 3043-3054.

[29] Juang B H, Hou W, Lee C H. Minimum classification error rate methods for speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 1997, 5(3): 257-265.

[30] Lei X, Senior A W, Gruenstein A, et al. Accurate and compact large vocabulary speech recognition on mobile devices[C]// Proc. Interspeech, 2013: 662-665.

[31] Kamvar M, Baluja S. A large scale study of wireless search behavior: Google mobile search[C]// Proc. SIGCHI conference on Human Factors in computing systems. ACM, 2006: 701-709.

[32] Kamvar M, Baluja S. Deciphering trends in mobile search[J]. Computer, 2007 (8): 58-62.

[33] Ling Z H, Qin L, Lu H, et al. The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007[C]// Proc. Blizzard Challenge Workshop. 2007.

[34] 李洪波, 于洪志. 基于藏语语音学知识的语音端点检测研究[C]// 中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集. 2007.

[35] 杨博,于洪志,关白. 语音合成技术在藏语 TTS 中的应用研究[J]. 西北民族大学学报(自然科学版), 2006, 01: 40-42+58.

[36] 孔江平. 藏语（拉萨话）声调感知研究[J]. 民族语文, 1995, 03: 56-64

[37] 鲍怀翘,徐昂,陈嘉猷. 藏语拉萨话语音声学参数数据库[J]. 民族语文, 1992,05: 10-20+9.

[38] 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.

[39] Yoshua B, Aaron C, Pascal V. Representation learning: a review and new perspectives.[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(8):1798-1828.

[40] 余凯,贾磊,陈雨强,徐伟. 深度学习的昨天、今天和明天[J]. 计算机研与发展, 2013, 09: 1799-1804.

[41] Yu D, Deng L. Deep learning and its applications to signal and information processing[J]. IEEE Signal Processing Magazine, 2011, 28(1): 145-154.

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

[43] 孙志军,薛磊,许阳明,王正. 深度学习研究综述[J]. 计算机应用研究, 2012, 08: 2806-2810.

[44] 贾丽会,张修如. BP 算法分析与改进[J]. 计算机技术与发展, 2006, 10: 101-103+107.

[45] Bishop C M. Neural networks for pattern recognition[M]. Oxford university press, 1995.

[46] Graves A, Liwicki M, Bunke H, et al. Unconstrained on-line handwriting recognition with recurrent neural networks[C]// Proc. Advances in Neural Information Processing Systems. 2008: 577-584.

[47] Graves A. Supervised sequence labelling[M]. Springer Berlin Heidelberg, 2012.

[48] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.

[49] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471.

[50] Schaefer A M, Udluft S, Zimmermann H G. Learning long-term dependencies with recurrent neural networks[J]. Neurocomputing, 2008, 71(13): 2481-2488.

[51] Schmidhuber J. Learning complex, extended sequences using the principle of history compression[J]. Neural Computation, 1992, 4(2): 234-242.

[52] Devert A, Bredeche N, Schoenauer M. Unsupervised Learning of Echo State Networks: A Case Study in Artificial Embryogeny[J]. Lecture Notes in Computer Science, 2008, 4926:278-290.

[53] H. W. Singer. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies[M]// Wiley-IEEE Press, 2009:237-243.

[54] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602-610.

[55] Wöllmer M, Schuller B, Eyben F, et al. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening[J]. IEEE Selected Topics in Signal Processing, 2010, 4(5): 867-881.

[56] Graves A, Liwicki M, Bunke H, et al. Unconstrained on-line handwriting recognition with recurrent neural networks[C]// Proc. Advances in Neural Information Processing Systems. 2008: 577-584.

[57] Wöllmer M, Blaschke C, Schindl T, et al. Online driver distraction detection using long short-term memory[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(2): 574-582.

[58] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. Neural

Computation, 2000, 12(10): 2451-2471.

[59] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[60]刘博,杨鸿武,甘振业,郭威彤. 利用SAMPA实现藏语的字音转换[J]. 计算机工程与应用, 2011, 35: 117-121.

攻读学位期间的研究成果

参与的科研项目：

[1]“藏汉双语跨语言语音合成的研究”, 甘肃省杰出青年基金计划项目(1210RJDA007), 2013-2015.

[2] “汉藏双语个性化语音合成中的语言建模的研究”, 国家自然科学基金项目(61263036), 2012-2016.

致谢

本论文是在导师杨鸿武教授的细心指导下完成的，在论文完成之际，衷心的感谢杨老师的严格要求和耐心指导。研究生期间，在杨老师带领的语音实验室团队中学会了很多，也收获了很多，杨老师对学术严谨的态度深深影响着我，杨老师学识渊博，知识广泛，在学术上不断探索，对每个细节都严抓不放，这种严谨的做事和科研态度使我受益匪浅。除此之外，杨老师随和、谦逊的做人态度也对我有很大的影响。在这里，我衷心的感谢杨老师为大家提供的良好的实验室学习环境，以及为大家耐心指导科研问题，特此向恩师表示崇高的敬意和衷心的感谢！

感谢物理与电子工程学院电子系的各位老师，是他们将渊博的知识细心的传授于我，让我学到了完善的理论知识，并对研究生所学的课程有了深刻的认识。感谢实验室已毕业的师兄师姐，借助于他们之前的研究使我对自己所研究的东西有了很好的认识。感谢实验室一起学习奋斗的同窗们。

我衷心感谢我的父母与我的亲戚朋友，感谢他们对我的支持和鼓励！最后，再次感谢所有对我有过帮助和支持的老师，同学和亲友们！