藏语声学建模方法研究

A Study on Subspace Clustering Algorithm for High-dimensional Data

专　　业：计算机科学与技术
学生姓名：李 健
指导教师：　教授

天津大学计算机科学与技术学院
二〇一六年十一月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**天津大学**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。

学位论文作者签名:　　签字日期:　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解**天津大学**有关保留、使用学位论文的规定。特授权**天津大学**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，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名:　　导师签名:

签字日期:　　年　　月　　日　签字日期:　　年　　月　　日

摘　　要

中文摘要应将学位论文的内容要点简短明了地表达出来，约500 800字左右（限一页），字体为宋体小四号。内容应包括工作目的、研究方法、成果和结论。要突出本论文的创新点，语言力求精炼。为了便于文献检索，应在本页下方另起一行注明论文的关键词（3-7个）。

关键词：　关键词1　关键词2　关键词3　关键词4

ABSTRACT

Externally pressurized gas bearing has been widely used in the field of aviation, semiconductor, weave, and measurement apparatus because of its advantage of high accuracy, little friction, low heat distortion, long life-span, and no pollution. In this thesis, based on the domestic and overseas researching……

Key words： Key Word 1, Key Word 2, Key Word, Key Word 4

目　录

第一章　绪论 · 2

1.1　自动语音识别简介 · · · · · · · · · · · · · · · · · · 2

1.2　本文研究内容及各章节安排 · · · · · · · · · · · · · · 3

第二章　背景介绍 · 5

2.1　自动语音识别 · 5

2.1.1　特征提取 · 6

2.1.2　声学模型 · 7

2.1.3　语言模型 · 9

2.1.4　发音词典 · 10

2.1.5　解码器 · 10

2.1.6　评价指标 · 10

2.2　Kaldi工具箱 · 10

第三章　声学模型训练 · · · · · · · · · · · · · · · · · · · 12

3.1　传统的GMM-HMM方法 · · · · · · · · · · · · · · · · 12

3.2　目前广泛使用的DNN-HMM方法 · · · · · · · · · · · · 12

3.3　LSTM · 12

第四章　拉萨方言语音识别 · · · · · · · · · · · · · · · · · 14

4.1　拉萨方言语音识别研究现状 · · · · · · · · · · · · · · 14

4.2　拉萨方言数据库及发音字典 · · · · · · · · · · · · · · 15

4.3　拉萨方言语音识别基准系统 · · · · · · · · · · · · · · 15

4.3.1　CD-DNN-HMM · · · · · · · · · · · · · · · · · · 16

4.3.2　声调相关特征 · · · · · · · · · · · · · · · · · · 16

4.3.3 模型融合 · · · · · 16
4.4 识别结果及分析 · · · · · 16
第五章 总结与展望 · · · · · 18
参考文献 · · · · · 19

第一章 绪论

1.1 自动语音识别简介

语音交互是人类社会最直接、最自然的沟通交流方式，而机器作为辅助人类生产及日常生活的工具，目前人类与各种机器交互的方式更多的还是依赖于键盘、鼠标、显示器等输入输出设备。如何摆脱鼠标、键盘，使得人与机器之间的沟通像人与人之间的沟通那样自然，是智能时代人类面临的重大挑战。想要实现人机对话，需要涉及语音识别、自然语言理解、语音合成等关键技术。其中，语音识别作为关键部分之一，就像机器的耳朵一样，机器需要依靠它来辨别人类到底在说些什么。

近些年来，随着科技的发展，在安静环境下使用近距离麦克风的应用场合，语音识别已达到实用阶段。纵观语音识别的发展史，20世纪50年代，贝尔实验室成功研制出世界上第一个语音识别系统Audrey，方法是基于元音的共振峰的测量，虽然该系统为针对特定说话人的孤立词识别，且只能识别十个英文数字的发音，但这意味着语音识别的时代开启了。20世纪60年代至70年代之间，语音识别领域取得了突破性进展。线性预测编码（Linear Predictive Coding）被应用在声学特征的提取上；动态时间规整（Dynamic Time Warping）技术用来解决模板匹配时非线性时间对齐的问题。这些关键性的突破使得特定说话人的孤立词识别成为可能。20世纪80年代，语音识别的任务开始从孤立词识别转向连续语音识别，比如识别连续朗读的数字串等。这一时期的重大进展是语音识别方法从模板匹配转为基于统计模型方法，其中最突出的是隐马尔科夫模型（Hidden Markov Model），该模型基于马尔科夫假设，实现了对时间序列结构的建模。该方法从80年代中期开始逐渐被世界范围内的研究机构广泛接受并成为主流的语音识别方法，直到今天，很多成熟的大规模连续语音识别系统依然没有脱离HMM的方法框架。20世纪90年代出现了很多判别训练方法，包括最小识别误差MCE(Minimum Classification Error)和最大互信息MMI(Maximum Mutual Information)等。相比于最大似然估计的训练方法，这些判别训练方法能够提供更好的识别性能。根据哥大ASR Lecture1 37页补充内容自2006年Hinton等人提出有效的训练深度神经网络算法开始，深度学习技术逐渐流行并在多个领域取得显著成果。在语音识别领域，深度学习用来进

行声学模型建模并获得巨大成功，尤其是对于大规模的识别任务而言。这要得益于反向传播算法的使用，以及越来越多的计算资源和训练数据。

1.2 本文研究内容及各章节安排

自20世纪60年代开始，近60年的技术积累使得语音识别性能已达到实用阶段，在某些特定的语音识别任务上，机器甚至已经超过人类。尤其是近几年深度神经网络取代传统的GMM（Gaussian Mixture Model）模型，使得识别率得到历史性突破。然而，这些性能上的突破大多都是针对英语、普通话等语料充足且已经被广泛研究理解的语种。对于许多语料匮乏语种来说，语言识别还停留在很初级的阶段。比如以藏语拉萨方言为例，目前还没有公开的比较成熟的语料库，同时也缺乏相应的拉萨方言的语音学知识，且由于藏语本身语言特性复杂，训练一个实用的语言模型十分困难，这些问题导致现阶段几乎还没有实用的藏语大规模连续语音识别系统。目前现有的关于藏语语音识别的研究主要集中在特征提取以及使用动态贝叶斯网络构建声学模型上。应用深度学习技术来对藏语声学模型建模的研究还非常少，藏语识别的研究也处在刚刚起步的阶段。即使是用于训练藏语声学模型的音素集合都还没有一个统一的参考标准。本研究从录制拉萨方言平衡语料库开始，设计了拉萨方言发音字典，尝试使用GMM-HMM、DNN-HMM及Tandem等方法训练了声学模型，通过爬取网络上的藏语文本数据训练得到语言模型，搭建了离线的拉萨方言语音识别系统。并且首次探索了如何利用拉萨方言的声调信息提高识别准确率。

本论文的章节安排如下：第一张为绪论部分，简要介绍了语音识别的任务及发展史；第二章为背景介绍，主要讲述了语音识别系统的各个组成部分及评价指标，并且介绍了本工作涉及到的语音识别工具箱。第三章总结了训练声学模型的各种方法，包括传统的GMM-HMM、目前广泛使用的DNN-HMM方法以及逐渐兴起的LSTM方法；第四章详细描述了搭建拉萨方言语音识别系统的过程，包括声学模型和语言模型的训练，以及声调特征提取的相关实验；第五章为总结和展望。

第二章　　背景介绍

本章主要介绍语音识别的一些基本概念，包括前端的特征提取、声学模型、语言模型、解码器、性能评价指标，另外简单介绍了本研究涉及到的语音识别工具箱——Kaldi。

2.1　自动语音识别

语音识别，顾名思义，是要把人的声音转化成文本，目标是在给定声音的前提下找到最有可能的文本序列。语音识别系统的结构可以用下图表示：

补充语音识别系统结构图

从图中可以看出，语音识别系统总共包含五个部分，输入的音频首先经过语音识别前端的特征提取部分，现假设语音长度为T，那么经过特征提取会得到一系列固定长度的频谱特征向量X_t，t=1,2,...T。语音识别的输出是一连串字符ω_k，k=1,2,...K。ω是系统认为最能匹配输入音频的文字序列。因此语音识别的目标可以表示为找到对应的ω使之满足

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmax}} \frac{P(X|\omega) * P(\omega)}{P(X)} \tag{2-1}$$

$P(\omega|X)$ 很难直接计算，根据贝叶斯公式

$$P(\omega|X) = \frac{P(X|\omega) * P(\omega)}{P(X)} \tag{2-2}$$

对于给定的输入X，P(X)对所有ω均为定值，因此

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmax}} P(X|\omega) * P(\omega) \tag{2-3}$$

其中，$P(X|\omega)$代表声学模型，$P(\omega)$代表语言模型。解码器根据声学模型和语言模型对ω的评分，搜索所有可能的ω，得到最优解。以下是语音识别系统各部分的详细介绍。

2.1.1 特征提取

对于语音识别系统而言，声学信号作为用户的唯一输入，需要承载用于识别的所有信息。如果仅对声学信号在时域上的波形进行分析，很难从中提取出对识别有用的特征，因为即使同一个人说同样一段话，单从波形上看都会有很大差别。然而，一个受过训练的人，可以通过语谱图区分不同的元音，因为元音的频率成分相对固定，不同元音的频谱图会有明显的差别。根据这一特性，我们可以在频域上对信号进行分析，从声学信号中提取与频率有关的特征，用来作为识别系统的输入。目前语音识别系统常用的声学特征包括：梅尔频率倒谱系数（MFCC）、感知线性预测(PLP)、Filter-bank等。下面以语音识别中常用的MFCC参数为例，详细介绍声学参数的提取过程。

梅尔频率倒谱系数(MFCC)作为语音识别中比较常用的声学特征参数，其原理是模仿人耳的听觉机理，将以赫兹为单位的频率变换成梅尔频率，使用在梅尔刻度上等距分布的梅尔滤波器组搜集不同频段的能量，通过逆离散傅里叶变换（IDFT）计算倒谱系数，实现声源和滤波器的分离，并降低不同维度特征之间的相关性。最后加入能量以及帧与帧之间的变换的信息。计算梅尔频率倒谱系数的详细过程如下：模拟信号经过采样和量化，转换为数字信号x[n]，n对应采样时刻。接下来对x[n]加窗，由于声学特征是用来区分语音信号中不同音素的，所以我们需要分析大致对应每个音素长度的部分的波形，这就需要对整个信号做加窗处理，窗口外部的信号全部设为零，只保留窗口内部的信号。一般情况下窗长设为25ms，每10ms 向前移动一个时间窗。这样每段音频都转化成了相互之间有重叠部分的固定长度的数字向量。MFCC提取过程中普遍使用的窗函数为Hamming Window，其公式为:

$$f(x)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{L}) & 0\le x\le L-1\\ 0 & \text{otherwise.}\end{cases} \tag{2-4}$$

接下来是对加窗后的数字信号做离散傅里叶变换，离散傅里叶变换的目的是计算信号在不同频段所包含的能量。其公式为:

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi}{N}kn} \tag{2-5}$$

DFT的N个输出对应N个离散频带，X[k]为复数，代表当前频率成分的幅值和相位。FFT是实现离散傅里叶变换的高效算法，但限制是N必须为2的正整数次幂。最后FFT 的输出表示每个频段的能量。前文已经提到过，MFCC是基于人耳的听

觉感知机理设计的，人耳对频率的感知是非线性的，超过1000Hz时，人类对频率的变换越来越不敏感。研究表明，在特征提取时，通过建模人类听觉的这种特性，能够提高识别系统的性能[Davis and Mermelstein]。MFCC通过引入梅尔刻度来模拟人耳的机制，梅尔刻度[Stevents and Volkmann]是描述声调的单位，两组信号如果在声调的感知上是等距的，那么它们的梅尔频率也是等距的。以赫兹为单位的频率与梅尔频率之间的关系可以使用如下公式表示：

$$mel(f) = 1127\ln(1 + \frac{f}{700}) \tag{2-6}$$

计算MFCC时，通过放置一组三角滤波器来收集不同频带的能量，三角滤波器在Mel刻度下是等宽均匀排布在整个频率范围内的。接着对每个梅尔滤波器的输出取对数，这样可以减弱特征对输入变化的敏感性，比如发音人与麦克风之间距离的变化。如果直接将梅尔滤波器输出取对数后的值作为特征，因为各个滤波器输出值之间相关性不为零，会造成后续训练高斯混合模型时协方差矩阵无法使用对角阵，所以需要进一步处理。倒谱（cepstrum）是对频谱取对数之后做逆傅里叶变换，倒谱可以实现声源和滤波器分离，并去除特征不同维度之间的相关性，因此取倒谱系数的前12维作为MFCC的特征。由于每一帧的能量和当前帧所属音素有关，可以将能量作为MFCC的一个维度：

$$E(m) = \sum_{t=t_1}^{t_2} x^2[t] \tag{2-7}$$

其中m表示帧的标号，t_1和t_2分别代表帧的起始时刻和终止时刻。除了能量之外，前后帧之间的变化信息也有助于识别不同的音素，所以MFCC一般还会加入倒谱系数每一维的一阶差分和二阶差分，以及能量的一阶差分和二阶差分。最终，从每一帧信号中提取出39维的MFCC向量，用于训练声学模型和识别。

2.1.2 声学模型

在隐马尔科夫模型（HMM）被用到语音识别之前，人们使用Dynamic Time Warping(DTW)来搭建最简单的语音识别器。DTW用来计算测试音频和模板音频之间的距离，假设需要识别0-9十个数字，那么就需要有0-9这十个数字分别对应的模板音频。当给定一个需要识别的音频时，使用DTW计算这个音频与10个模板音频的距离，从中挑选距离最短的一个，对应的模板的数字就是识别结果。由于语音本身的时变特性，不可能要求识别的音频和模板音频时长保持一致，对于时长不匹配的问题，DTW通过引入warping function Tx(t), Ty(t),t=1,2,...T.对测试音频和模板音频进行非线性对齐。给定样本X和Y，有很多可选的对齐方式，

随着Tx和Ty 的增加，可选的对齐方式数量成指数级增长。DTW使用Dynamic Programming（DP），极大地缩小搜索空间，通过保存历史信息，找到最优路径。但DTW方法可扩展性比较差，当语音识别的词汇量增加时，模板的数量也就随之增加，当识别任务变为大规模语音识别时，DTW计算测试音频与每个模板之间的距离已经变得不切实际。由此，隐马尔科夫模型替代了DTW，一直沿用至今。

隐马尔科夫模型是概率模型，用来描述由马尔科夫链随机生成不可观测的状态序列的过程。状态序列里的每个状态都会在当前时刻t生成一个观测，由此组成观测序列。隐马尔可夫模型遵从两个基本假设：1）隐马尔可夫模型在任意t时刻的状态只与前一时刻的状态有关，与其它任意时刻的状态及观测都无关；2）任意时刻的观测只依赖于该时刻的状态，与其它状态及观测无关。基于以上的假设，隐马尔科夫模型可以由三个参数决定：1）状态转移概率；2）观测概率；3）初始状态概率。给定这三个参数，隐马尔科夫模型就定固定了。以下是隐马尔科夫模型三个参数的定义：设S是所有可能的状态的集合，M是所有可能的观测的集合。

$$S = s_1, s_2, ..., s_H, \quad M = m_1, m_2, ..., m_K$$

，其中，H为所有可能的状态数，K为所有可能的观测数。现有长度为T的状态序列Θ，以及状态序列对应的观测序列Φ。

$$\Theta = (\theta_1, \theta_2, ..., \theta_T), \quad \Phi = (\phi_1, \phi_2, ..., \phi_N)$$

则状态转移概率矩阵:

$$A = [a_{ij}]_{H \times H} \tag{2-8}$$

$$a_{ij} = P(\theta_{t+1} = s_j | \theta_t = s_i), i = 1, 2, ..., H; j = 1, 2, ..., H \tag{2-9}$$

a_{ij}代表t时刻的状态s_i在t+1时刻跳转到s_j的概率。观测概率矩阵可以用B来表示:

$$B = [b_j(n)]_{H \times K} \tag{2-10}$$

$$b_j(n) = P(\phi_t = m_n | \theta_t = s_j), n = 1, 2, ..., K; j = 1, 2, ..., H \tag{2-11}$$

$b_j(n)$代表t时刻处于状态j的前提下，观测到ϕ_n的概率。初始状态概率向量可以用Ψ表示：

$$\Psi = (\psi_1, \psi_2, ..., \psi_i, ..., \psi_H) \tag{2-12}$$

$$\psi_i = P(\theta_1 = s_i), \quad i = 1, 2, ..., H \tag{2-13}$$

ψ_i代表初始时刻处于状态s_i的概率。

隐马尔科夫模型可以由以上定义的状态转移概率矩阵A，观测概率矩阵B，以及初始状态概率向量Ψ完全确定。至此，隐马尔科夫模型可以表示为$\Pi = (A, B, \Psi)$。定义完模型之后，接下来引出隐马尔科夫模型的三个基本问题：

1) 概率计算问题。给定模型$\Pi = (A, B, \Psi)$及观测序列$\Phi = (\phi_1, \phi_2, ..., \phi_N)$，计算在模型Π下观测到Φ这一观测序列的概率；

2) 解码问题。给定模型$\Pi = (A, B, \Psi)$及观测序列$\Phi = (\phi_1, \phi_2, ..., \phi_N)$，求解最有可能的状态序列Θ；

3) 模型参数训练问题。已知观测序列Φ，计算模型参数(A, B, Ψ)，使得似然概率$P(\Phi|\Pi)$最大。

对应到语音识别中，隐马尔科夫模型被用来对连续的特征序列建模。假设现在有一个英文数字0-9的识别任务，如何利用隐马尔科夫模型识别数字呢？根据2.1.1的MFCC特征提取过程，一个1s的音频，如果每10ms移动一帧，那么会得到将近100个39维的MFCC特征向量。这些特征向量对应的就是观测序列$\Phi = (\phi_1, \phi_2, ..., \phi_N)$。在语音识别中，首先会给声音的最小单位建立隐马尔科夫模型，对于英语来说，最小单位就是音素。而词表中的每个单词都能由音素构成。比如‘one’这个单词，可以用W, AH, N这三个音素来表示其发音。对应音素的隐马尔科夫模型包含三个状态，分别对应发音的起始阶段、中间阶段和结束过程。表示单词的隐马尔科夫模型就是用组成其发音的音素的隐马尔科夫模型串联起来的。下图表示了one这个单词对应的隐马尔科夫模型，其它9个数字对应的隐马尔科夫模型同理。插入隐马尔科夫模型图片。有了词表里所有单词对应的隐马尔科夫模型$\Pi_i, i = 0, 1, ...9$，同时也有了观测序列$\Phi = (\phi_1, \phi_2, ..., \phi_N)$，那么如何识别当前音频到底是哪一个数字呢？这就对应到了隐马尔科夫模型的第一个基本问题，概率计算问题。我们需要分别计算$P(\Phi|\Pi_i), i = 0, 1, ..., 9$这10个似然概率。其中似然概率最大的模型对应的数字就是识别结果。

$$result = \arg\max_i P(\Phi|\Pi_i) \tag{2-14}$$

2.1.3 语言模型

根据公式2-3，可以看到目标函数包含$P(\omega)$这一项。ω表示单词的序列，$P(\omega)$代表这一单词序列的先验概率。为什么要加入这一项先验概率呢？根据2.1.2，我们已经知道声学模型可以计算似然概率$P(O|\Omega)$，其中O代表观测

向量，Ω代表单词序列。但是如果仅凭声学模型得到的结果去判断最可能的单词序列Ω的话，没办法处理语言中同音字的情况。比如英语里的'good four me'和'good for me'，这两组句子从发音上将几乎没有差别，因此通过声学模型得到的结果也几乎相同。但是，'good for me'更符合英语的规则和使用习惯，$P('good\quad for\quad me')$要明显大于$P('good\quad four\quad me')$。因此，在目标函数中加入语言规则的先验知识，能更好地区分从声学角度上比较容易混淆的语音。然而，语言的规则并不是通过显式地规定语法规则，而是通过统计模型。比较常用的语言模型为N-gram模型。

2.1.4 发音词典

2.1.5 解码器

2.1.6 评价指标

2.2 Kaldi工具箱

第三章　　声学模型训练

3.1　传统的GMM-HMM方法

3.2　目前广泛使用的DNN-HMM方法

3.3　LSTM

第四章　拉萨方言语音识别

拉萨方言作为语料匮乏语言的一种，想训练其语音识别系统，难点在于收集足够多的带文本标签的语音数据，同时也缺少足够的藏语语音学和语言学的相关知识。这些限制条件使得拉萨方言语音识别的研究进展缓慢，远远落后于目前英语和汉语的语音识别技术水平。在本研究中，我们录制了一个小规模的拉萨方言数据库，从头开始搭建了拉萨方言离线识别系统。在语音数据不充足的条件下，调查了拉萨方言声调信息对语音识别性能的影响。由于拉萨方言目前还没有一个公认的声调系统，因此本论文中，我们采用了四个声调的声调模式，设计了带调的音素集合，并且在特征层面加入了声调相关的信息。实验结果表明，利用声调信息之后，识别的准确率相对提升16.0%。以下是拉萨方言语音识别系统实验的详细介绍。

4.1　拉萨方言语音识别研究现状

拉萨方言属于西藏中部方言的一种，使用者包括拉萨以及周边地区的居民。由于拉萨在政治和文化方面的重要性，拉萨方言相关的研究近些年开始受到越来越多的关注。现有的拉萨方言相关研究工作，其中一部分为拉萨方言声学模型的研究。[2]是较早的使用DTW做拉萨方言孤立词识别的，[3,4]是使用传统GMM做拉萨方言连续语音识别的。很明显，以上的研究所用技术已经落后于目前流行的方法。[5,6]关注的是如何使用神经网络的共享隐层解决拉萨方言训练数据不足的问题。虽然[5,6]等最新的研究使用了流行的深度学习技术，但是很少有相关工作提到如何使用拉萨方言本身的特性去提高识别性能。

拉萨方言属于单音节的有调语言，每个拉萨方言的单字都是一个带调的音节，声调在区分同音字上扮演着很重要的角色，尤其是在缺少较强的上下文信息的情况下。然而，很少有研究提到如何利用拉萨方言的声调去提高系统的识别性能。如果能对拉萨方言的声调建立准确模型，将会在很大程度上提高识别系统的准确度。

4.2 拉萨方言数据库及发音字典

藏语拉萨方言数据库由天津大学认知计算与应用重点实验室(CCA)与中国社会科学院民族学与人类学研究所合作录制，共包含13名男性发音人和10名女性发音人。发音人均是以拉萨方言为母语的中央民族大学本科生。每位发音人录制相同的3,100句藏语音素平衡句，句子的平均时长为3.2秒。录制环境为安静的办公室环境。音频信号的获取采用单声道、16KHz采样率、16bit量化，保存为.wav格式的音频文件。数据经过人工校对，剔除掉不合格的音频数据，最后可用的数据总时长为35.92小时。将数据库分为三个部分，其中训练集数据用作模型训练，包含7名女性发音人和10名男性发音人，共36,090个句子，总时长为31.9小时；测试集用作模型测试，包含3名女性发音人和3名男性发音人，共2,664个句子，总时长为2.41小时；开发集用于选择模型参数，其发音人和训练集发音人一致，共1700个句子，总时长为1.51小时。训练集和测试集发音人和发音句子没有交叉，保证了模型测试结果的准确性。

发音字典由合作单位中国社会科学院民族学与人类学研究所提供，字典采用声韵母组合的规则，共包含29个声母，48个韵母。字典条目为2,100。基本涵盖了所有藏语拉萨方言数据库中出现的藏文字。下表是发音字典所用声韵母集合：添加拉萨声韵母集合

4.3 拉萨方言语音识别基准系统

为了验证声调特征的有效性，我们首先搭建了基准系统。在基准系统中，我们尝试了两种不同的DNN声学模型框架，一种是DNN-HMM；另外一种是Tandem方法。其中DNN-HMM使用深度神经网络代替传统的GMM模型，用来计算上下文相关的状态发射概率，使用HMM对时序关系建模；而Tandem方法是使用带有bottle-neck层的神经网络做为特征提取的手段，将bottle-neck层的输出与传统的MFCC特征拼接在一起作为最后的特征，并用这些特征训练GMM-HMM模型。根据以往研究人员的经验，Tandem方法一般能够达到DNN-HMM方法的效果，不过要比DNN-HMM实现起来稍微繁琐复杂。在本论文中，我们在语料不是很充足的前提下，分别尝试了这两种方法，并对结果进行了分析比较。

4.3.1 CD-DNN-HMM

DNN-HMM方法框架属于深度神经网络在语音识别应用中的一种。DNN-HMM利用了DNN模型强大的表征能力，同时也保留了HMM模型对序列建模的能力。当DNN的输出单元对应的不是单音素模型(monophone)，而是捆绑后的三音子模型(senone)时，DNN-HMM就被称为是CD-DNN-HMM(context-dependent-DNN-HMM)。与传统的GMM-HMM模型相比，DNN-HMM使用强大的神经网络模型计算HMM对应的发射概率。很多研究表明，CD-DNN-HMM在很多大词汇量连续语音识别的任务上都要强过GMM-HMM。由于CD-DNN-HMM与GMM-HMM模型共享senones和HMM模型，所以第一步需要训练GMM-HMM，并以GMM-HMM为基础去训练CD-DNN-HMM。例如，训练集里每个句子的senones的切分信息都是由GMM-HMM模型生成的。

在本论文中，对于GMM-HMM系统，输入特征首先经过谱均值方差归一化(CMVN)处理，接下来使用处理后的特征训练单音素模型(monophone system)，再接着使用单音素模型和维特比算法，对训练数据里的每个句子做强制对齐，得到音素的切分信息，最后，使用得到的切分信息训练三音素模型(triphone system)。在训练三音素模型的过程中，使用了线性判别分析(LDA)和最大似然线性变换(MLLT)等特征变换方法提高模型的性能。

对于CD-DNN-HMM模型，其模型结构为六个隐层，每个隐层2048个节点。训练DNN的输入特征为当前帧加前后五帧，总共11帧的MFCC参数。DNN的初始化参数通过预训练受限玻尔兹曼机(restricted RBM)得到。在预训练步骤完成后，训练阶段采用随机梯度下降算法(SGD)，训练的标签信息通过GMM-HMM对训练句子做强制对齐得到。

4.3.2 声调相关特征

4.3.3 模型融合

4.4 识别结果及分析

fjsdl

第五章　　总结与展望

拉萨方言作为有调语言，声调对于区分同音字起到了关键的作用。但目前对于拉萨方言具体有几个调还存在争议，这对使用声调信息造成了困难。本研究通过调研相关文献，并结合已录制的拉萨方言音频数据库，采用四个声调的声调系统，包括高平调(55)、升调(13)、降调(51)、升降调(132)。最终确定了拉萨方言带调音素集合，共29个声母，48个韵母，其中每个韵母用四个不同的调来区分。在特征层面，提取每一帧的基频值，再结合MFCC参数，构成声调相关的声学特征参数。为了验证声调信息有助于提升拉萨方言的识别结果，本研究搭建了完整的识别系统，使用不同的音素集合和输入特征，分别训练三音素模型和DNN-HMM模型，得到字级别的识别结果。首先使用传统的39维MFCC特征作为输入特征，音素集合使用29个声母和48个韵母；之后，输入特征不变，音素集合中的韵母使用四个声调来区分；最后，输入特征加入每一帧的基频值，音素集合中的韵母使用四个声调来区分。实验的训练数据31.9小时，测试集2.41小时。对于DNN模型，使用区分声调的音素集合，字级别错误率相对下降6.1%，使用声调相关的输入特征，可以进一步降低识别错误率。最后，使用区分声调的音素集合，配合声调相关的输入特征，字级别错误率与基准系统相比相对下降11.1%。该研究验证了声调信息对拉萨方言语音识别的重要性。

参考文献

[1] 胡伟. LaTeX 2ε 完全学习手册 [M]. 北京: 清华大学出版社, 书号: 978-7-302-24159-1, 2011.

[2] 邓建松, 彭冉冉, 陈长松. LaTeX 2ε 科技排版指南 [M]. 北京: 科学出版社, 书号: 7-03-009239-2/TP.1516, 2001.

[3] Lamport L. LaTeX — A Document Preparation System: User's Guide and Reference Manual [M]. 2nd ed. Reading, Massachusetts: Addison-Wesley, 1985.

[4] Knuth D E. The TeXbook [M]. Reading, Massachusetts: Addison-Wesley, 1986.

[5] Knuth D E. Computer Modern Typefaces [M]. Reading, Massachusetts: Addison-Wesley, 1986.

[6] Bezos J. The titlesec and titletoc Packages [M]. 2nd ed. Cityname: University of SomeName, 2002.

[7] P Oostrum, ifuleyou@bbsctexorg 译. LaTeX下的页面布局 [M]. 天津: 某某大学出版社, 2001.

[8] Shell M. How to Use the IEEEtran LaTeX Class [J]. Journal of LaTeX Class Files, 2002, 1 (11): 10–20.

[9] TeXGuru. LaTeX2ε用户手册 [M]. 天津: 某某大学出版社, 1999.

[10] K Reckdahl 原著, 王磊 译. Using Import graphics in LaTeX2ε, LaTeX2ε插图指南 [M]. 天津: 某某大学出版社, 2000.

[11] McDonnell J R, Wagen D. Evolving Recurrent Perceptions for Time-Series Modeling [J]. IEEE Trans. on Neural Networks, 1994, 5 (1): 24–38.

[12] XYao. Evolutionary Artifitial Neural Networks [J]. J. Of Neural Systems, 1993 (4): 203–222.

[13] 宋乐. 异源图像融合及其评价方法的研究 [D]. 天津: 天津大学, 2008.

[14] Agrawal A, Raskar R. Resolving objects at higher resolution from a single motion-blurred image [C]. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007: 1–8.

[15] Zhang J, Li X, Chen J, et al. A tree parent storage based on hashtable for XML construction [C]. In Communication Systems, Networks and Applications (ICCSNA), 2010 Second International Conference on, 2010: 325–328.

[16] SNiwa, Suzuki M, Kimura K. Electrical Shock Absorber for Docking System Space [C]. In IEEE International Workshop on Intelligent Motion Control, Istenbul, 1990: 825–830.