

# 拉萨方言语音识别声学建模方法研究

## A Study on Acoustic Modeling for Lhasa Dialect Speech Recognition

专 业: 计算机科学与技术  
学生姓名: 李 健  
指导教师: 李雪威 副教授

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

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，建立了拉萨方言带调音素集合。在特征层面，尝试了使用两种不同的基频提取方法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用来训练声学模型。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，训练DNN-HMM模型，得到字级别的识别结果。实验的训练数据31.9小时，测试集2.41小时。识别结果表明，对于DNN模型，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。

关键词： 语音识别 拉萨方言 声学模型 声调

# **ABSTRACT**

As a tonal language, tone of Lhasa dialect is essential to help discriminate homophones. However, it remains controversial how many tones does Lhasa dialect have. This uncertainty brings difficulty to utilize tonal information in ASR of Lhasa dialect. In this study, we created a small-scale corpus and built the ASR system for Lhasa dialect from scratch. We adopted a four-tone pattern and designed a phone set based on the four contour contrasts scheme. After comparing different acoustic modeling methods, we use DNN model to investigate influences of the tonal information on the performance of Lhasa Tibetan speech recognition. The experimental results showed that the character error rate (CER) was improved 11% after applying the tone based phone set and pitch-related features to DNN-HMM based speech recognition by comparing to that without tonal information. This preliminary study revealed that the tonal information plays an important role in speech recognition of Tibetan Lhasa dialect.

**Key words:** ASR, Lhasa dialect, Acoustic model, tone

# 目 录

第一章 绪论 ······	1
1.1 自动语音识别简介 ······	1
1.2 本文研究内容及各章节安排 ······	2
第二章 背景介绍 ······	3
2.1 自动语音识别 ······	3
2.1.1 特征提取 ······	4
2.1.2 声学模型 ······	6
2.1.3 语言模型 ······	8
2.1.4 基于WFST的解码图 ······	11
2.1.5 评价指标 ······	14
2.2 Kaldi工具箱 ······	14
第三章 声学模型训练 ······	17
3.1 传统的GMM-HMM方法 ······	17
3.1.1 GMM模型的定义 ······	17
3.1.2 GMM模型的参数估计 ······	18
3.1.3 GMM-HMM模型训练 ······	19
3.2 目前广泛使用的DNN-HMM方法 ······	19
3.2.1 DNN-HMM模型训练步骤 ······	20
第四章 拉萨方言语音识别 ······	22
4.1 拉萨方言语音识别研究现状 ······	22
4.2 拉萨方言数据库及发音字典 ······	23

4.3 拉萨方言语音识别基准系统	23
4.3.1 CD-DNN-HMM	25
4.3.2 Tandem	25
4.3.3 两种建模方法的音素级识别结果	26
4.3.4 CD-DNN-HMM基准系统	28
4.4 拉萨方言声调系统	30
4.4.1 拉萨方言的四个声调类型	31
4.5 声调特征提取	31
4.5.1 SAcC方法	33
4.5.2 Kaldi-Pitch方法	33
4.5.3 声调相关特征	34
4.6 加入声调信息的识别系统	35
4.6.1 系统融合	36
第五章 总结与展望	37
参考文献	38
发表论文和参加科研情况说明	40
致 谢	41

# 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。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，机器甚至已经超越人类。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<sup>[1]</sup>。

### 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图2-1表示：从

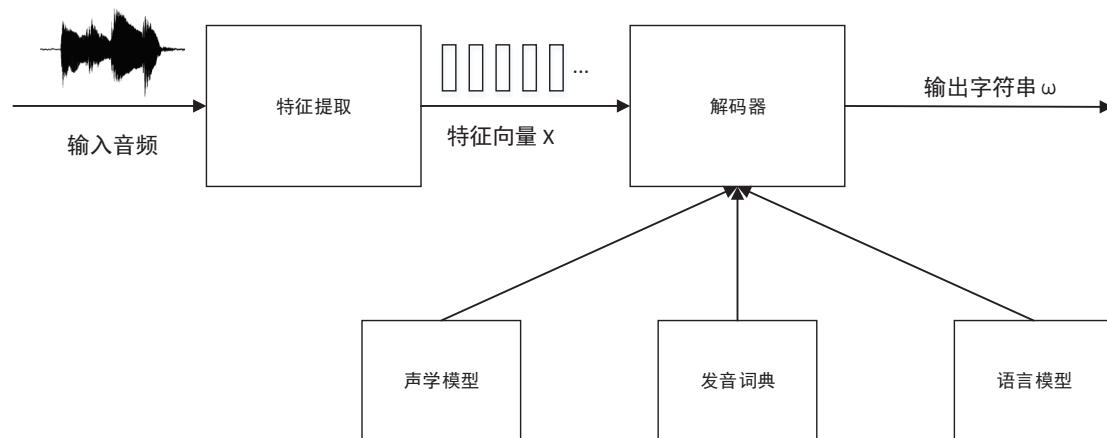


图 2-1 语音识别系统结构图

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  $X_t$ ,  $t=1,2,\dots,T$ 。语音识别的输出是一连串字符  $\omega_k$ ,  $k=1,2,\dots,K$ 。 $\omega$  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  $\omega$  使之满足

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

$P(\omega|X)$  很难直接计算，根据贝叶斯公式

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

对于给定的输入X， $P(X)$ 对所有 $\omega$ 均为定值，因此

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

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 $\omega$ 的评分，搜索所有可能的 $\omega$ ，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）<sup>[2]</sup>、感知线性预测(PLP)<sup>[3]</sup>、Filter-bank<sup>[4]</sup>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转换成梅尔频率，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：

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

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：

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

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<sup>[2]</sup>。MFCC通过引入梅尔刻度来模拟人耳的机制，梅尔刻度<sup>[5]</sup>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：

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

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：

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

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)<sup>[6]</sup>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  $T_x(t)$ ,  $T_y(t), t=1, 2, \dots, T$ 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 $T_x$ 和 $T_y$ 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。

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

，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 $\Theta$ ，以及状态序列对应的观测序列 $\Phi$ 。

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

则状态转移概率矩阵:

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

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

$a_{ij}$ 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:

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

$$b_j(n) = P(\phi_t = m_n | \theta_t = s_j), n = 1, 2, \dots, K; j = 1, 2, \dots, H \quad (2-11)$$

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

$$\Psi = (\psi_1, \psi_2, \dots, \psi_i, \dots, \psi_H) \quad (2-12)$$

$$\psi_i = P(\theta_1 = s_i), \quad i = 1, 2, \dots, H \quad (2-13)$$

$\psi_i$ 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 $\Psi$ 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 $\Phi = (\phi_1, \phi_2, \dots, \phi_N)$ ，计算在模型 $\Pi$ 下观测到 $\Phi$ 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 $\Phi = (\phi_1, \phi_2, \dots, \phi_N)$ ，求解最有可能的状态序列 $\Theta$ ；
- 3) 模型参数训练问题。已知观测序列 $\Phi$ ，计算模型参数 $(A, B, \Psi)$ ，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 $\Phi = (\phi_1, \phi_2, \dots, \phi_N)$ 。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。图 2-2 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。

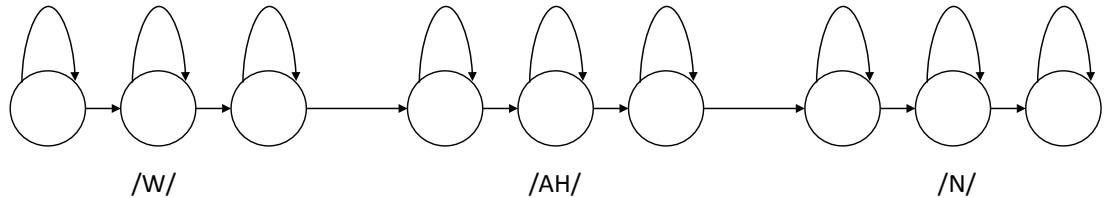


图 2-2 单词'one'对应的HMM模型

有了词表里所有单词对应的隐马尔科夫模型  $\Pi_i, i = 0, 1, \dots, 9$ ，同时也有了观测序列  $\Phi = (\phi_1, \phi_2, \dots, \phi_N)$ ，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  $P(\Phi|\Pi_i), i = 0, 1, \dots, 9$  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。

$$\text{result} = \underset{i}{\operatorname{argmax}} P(\Phi|\Pi_i) \quad (2-14)$$

图 2-3 为识别 0-9 十个数字时整体的隐马尔科夫模型框架。以上均为语音识别的解码过程，有关具体的模型参数训练问题，将在 3.1 中详细介绍。

### 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。 $\omega$  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， $\Omega$  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  $\Omega$  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， $P('good for me')$  要明显大于  $P('good four me')$ 。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。

N-gram 模型是一种统计模型，由于其简单有效，在语音识别中得到了广泛的应用。语言模型的目标为计算字符串  $\omega$  作为一个句子出现的概率  $P(\omega)$ 。对于一

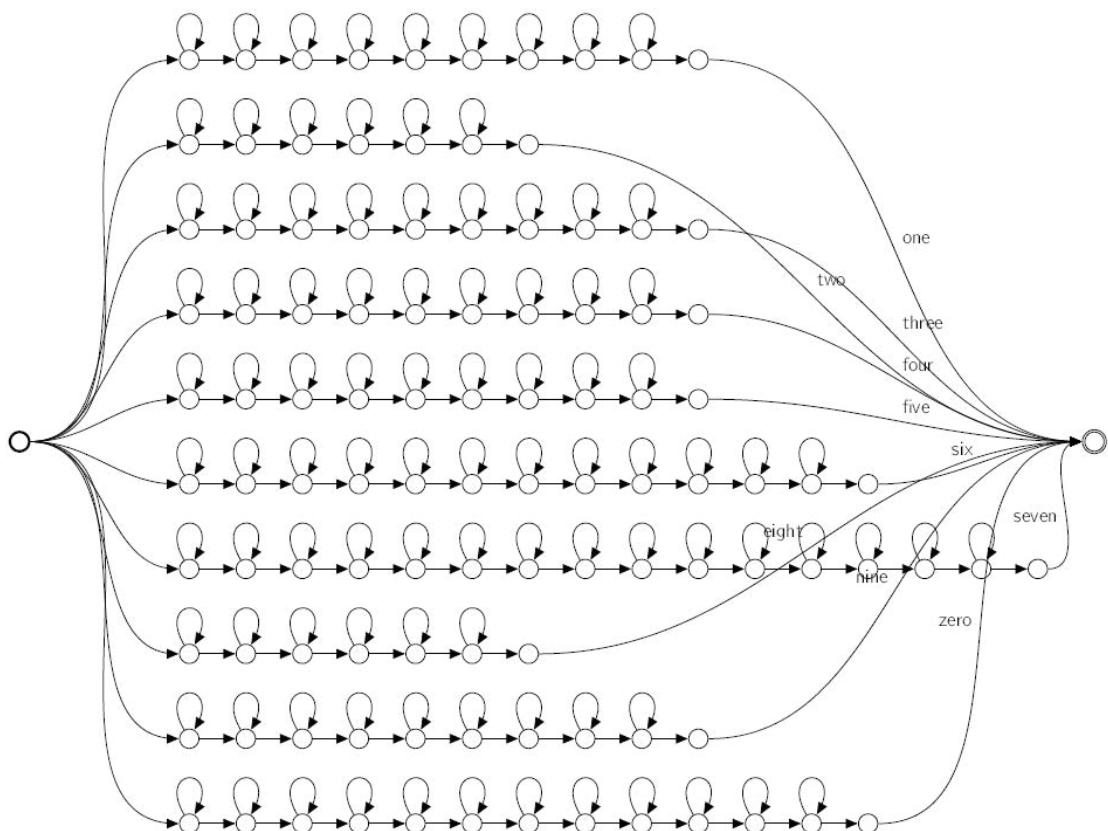


图 2-3 识别0-9十个数字的隐马尔科夫模型

个由 $m$ 个字符组成的序列 $\omega = c_1c_2...c_m$ 而言，

$$P(\omega) = P(c_1)P(c_2|c_1)P(c_3|c_1c_2)...P(c_m|c_1c_2...c_{m-1}) \quad (2-15)$$

$$= \prod_{i=1}^m P(c_i|c_1...c_{i-1}) \quad (2-16)$$

上式中，产生第 $i$ 个字符 $c_i$ 的概率是由已经产生的 $i-1$ 个字符 $c_1c_2...c_{i-1}$ 决定的。字符序列 $c_1c_2...c_{i-1}$ 可以看成是 $c_i$ 的历史信息。对于这种计算方法，假设词表长度为 $N$ ，那么 $c_1c_2...c_{i-1}$ 将会有 $N^{i-1}$ 种可能，当 $i$ 增加到一定程度，几乎已经无法计算 $P(c_i|c_1...c_{i-1})$ 。因此，需要对模型做出一定的简化。 $N$ -gram模型假设，对于预测当前字符 $c_i$ 产生的概率，只需要参考前 $N-1$ 个已经出现的字符 $c_{i-n+1}c_{i-n+2}, c_{i-1}$ ，而不需要考虑所有的历史信息。这样一来， $P(\omega)$ 的计算公式就变为：

$$P(\omega) = \prod_{i=1}^m P(c_i|c_{i-n+1}^{i-1}) \quad (2-17)$$

通常 $N$ 的取值不能太大，否则依旧会出现计算量过大导致参数无法估计的问题，所以通常选取 $N=1,2,3$ 。 $n=1$ 时，第 $i$ 个词出现的概率独立于前面的词，称为unigram； $n=2$ 时，第 $i$ 个词出现的概率只与前一个词 $c_{(i-1)}$ 有关，称为bigram； $n=3$ 的情况最多，即第 $n$ 个词出现的概率与前两个词 $c_{(i-1)}(i-2)$ 有关，记为trigram。因为句子中每个词出现的概率可以通过统计语料中该词出现的次数得到。以unigram为例，为了计算条件概率 $P(c_i|c_{i-1})$ ，可以通过统计二元词序列出现的次数来计算得到，计算公式如下：

$$P(c_i|c_{i-1}) = \frac{\text{count}(c_{i-1}c_i)}{\sum_{c_i} \text{count}(c_{i-1}c_i)} \quad (2-18)$$

在使用 $N$ -gram语言模型计算出现的概率时，由于训练数据稀疏，对于很多正常的句子 $\omega$ ，其出现的概率 $P(\omega) = 0$ 。这就意味着，在语音识别的应用中，即使句子 $\omega$ 在声学模型下的打分很高，但是由于 $P(\omega) = 0$ ，使得识别结果永远不会出现 $\omega$ 。解决这种错误的一种典型的方法为平滑方法，其基本思想为提高低概率，降低高概率，从而解决零概率的问题。实际应用中最简单的平滑方法之一为加法平滑，基本思想为：假设每一个 $n$ 元语法发生的次数比实际统计的次数多 $\delta$ 次， $0 \leq \delta \leq 1$ 。此时：

$$P_{add}(c_i|c_{i-n+1}^{i-1}) = \frac{\delta + \text{count}(c_{i-n+1}^i)}{\delta|V| + \sum_{c_i} \text{count}(c_{i-n+1}^i)} \quad (2-19)$$

其中， $|V|$ 表示词汇表单词的个数。此外，常用的平滑方法还有古德-图灵估计法，Katz平滑方法，Jelinek-Mercer平滑方法，绝对减值法，Kneser-Ney平滑方法

等<sup>[7]</sup>。

#### 2.1.4 基于WFST的解码图

对于大词汇量连续语音识别，需要加入语言模型来提高识别的准确性，减少由同音字或易混淆语境导致的识别错误。但如何才能将语言模型与声学模型配合起来呢？当词汇量很小时，比如0-9个数字的孤立词识别 2-3。语言模型一般用一元语法模型，先验概率  $P(\omega_i)$  可以直接填在每个单词的HMM模型的后面。但如果识别任务变为大词汇量连续语音识别时，语言模型一般采用二元或三元语法，此时就无法直接在每个单词的HMM模型上直接添加语言模型。并且，对于大词汇量任务来说，不同单词之间需要共用音素的隐马尔科夫模型，所以还需要考虑发音词典。同时，当识别任务为连续语音识别时，必须要考虑到协同发音的影响，同样一个音素，在不同的语境下，发音差别很大，这就需要区分不同上下文下的音素，一般使用三音素模型，即只考虑前一个和后一个音素对当前音素的影响。综上，对于大词汇量连续语音识别，需要加入语言模型、发音词典、音素的上下文关系。如何把所有这些音素同时考虑进去，构建一个大的隐马尔科夫模型呢？这就需要引入有限状态机的概念。

这里有两种类型的有限状态机，一种是有限状态接收器(Finite State Acceptor)。它包含一个起始状态，一个或多个终止状态，不同状态之间的跳转由一个带有向箭头的弧表示，弧上的符号代表状态跳转时，FSA的输出。图 2-4 为一个有限状态接收器的例子，图中弧上的符号  $\epsilon$  代表状态跳转时输出为空字符。

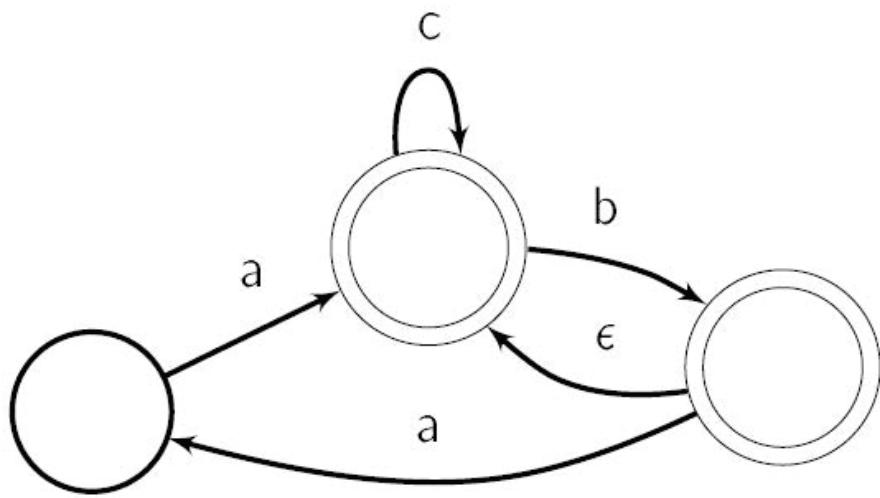


图 2-4 有限状态接收器(FSA)

每个FSA都会有对应的可接收字符串。FSA从初始状态跳转到终止状态，过

程中所有的输出按顺序组成的字符串都是可接收字符串。比如字符串'abb'是图 2-4 中的FSA的一个可接收字符串。

除了有限状态接收器，还有一种是有限状态转换器(Finite State Transducer)，FST与FSA的差别是FST每条弧上有两个符号，分别对应输入符号和输出符号。图 2-5 展示的是一个FST。

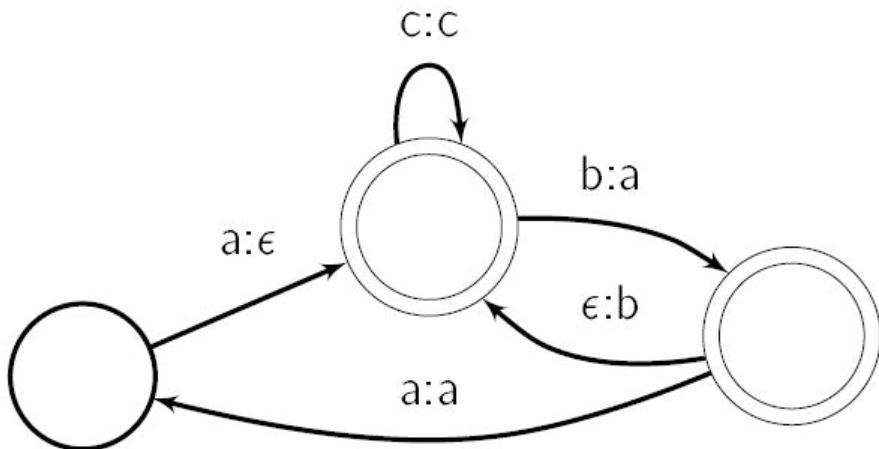


图 2-5 有限状态转换器(FST)

与FSA类型，FST也有接收状态。当FST由起始状态跳转到终止状态时，过程中每个弧的输入符号按顺序排列得到字符串s，每个弧的输出符号按顺序排列得到字符串t，此时成FST可接收(s,t)。

有限状态机接收器A与有限状态转换器T之间可进行组合操作，用符号 $AoT$ 表示。通过组合(Composition)操作，T可以替换A的每条弧上标签，得到一个新的有限状态接收器B，T上的每条弧上的输入符号和输出符号代表替换规则。新生成的B接收字符串t，当且仅当A接收字符串s,且T接收(s,t)。图 2-6 表示FST与FSA通过组合操作，生成新的FSA。

对于语音识别中用到的隐马尔科夫模型、发音词典以及语言模型，都可以用一个加权有限状态接收器(WFST)来表示。加权有限状态接收器与有限状态接收器的区别是在每个弧上增加了一个权重，在语音识别中，这个权重一般代表概率或者概率的对数。图 2-7 中，a、b、c分别为用WFSA表示的语言模型、发音词典、音素的隐马尔科夫模型。

有了加权有限状态接收器和加权有限状态转换器，可以通过组合操作，将语言模型、发音词典、音素的上下文信息、音素的隐马尔科夫模型整合到一起，最终形成一个整体的隐马尔科夫模型。具体操作是，首先将语言模型用加权有限状态接收器G表示，然后通过设计加权有限状态转换器，将语言模型里的每个

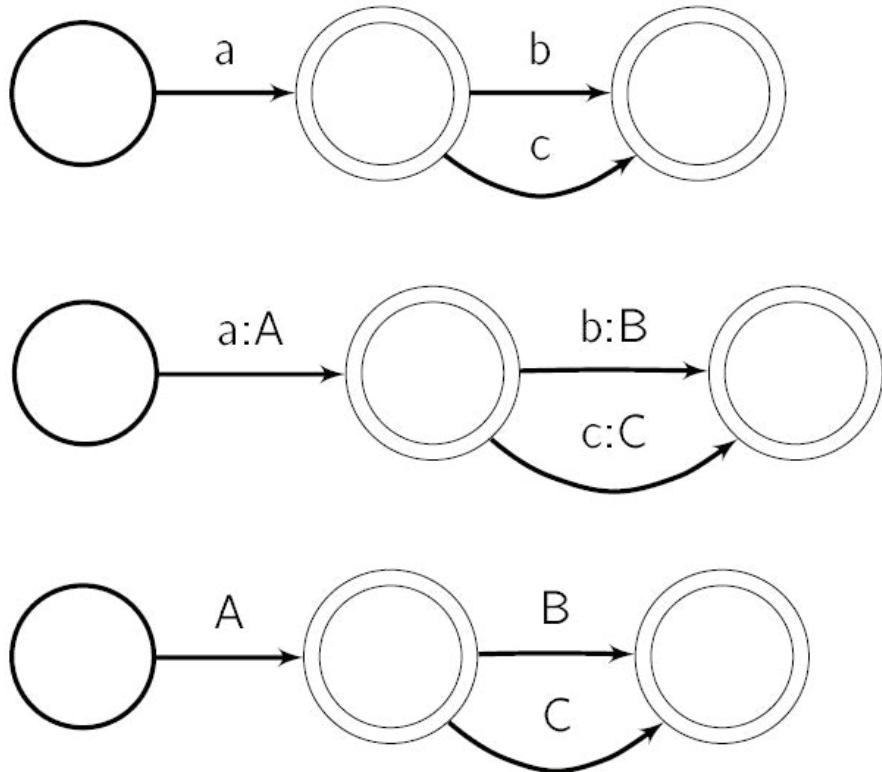
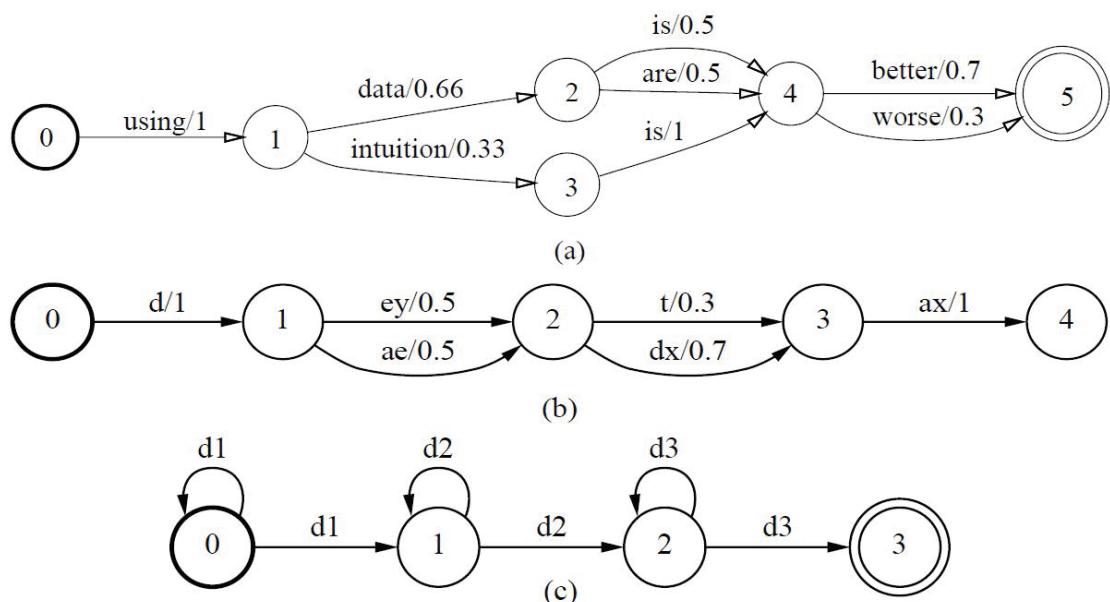


图 2-6 FST与FSA之间的组合操作

图 2-7 (a)、(b)、(c)分别为用WFSA表示的语言模型、发音词典、隐马尔科夫模型<sup>[8]</sup>

单词替换成其对应的发音组合，得到新的加权有限状态接收器P。接着再通过组合操作，将P中所有上下文无关音素变成其对应的上下文相关音素，得到新的加权有限状态接收器C。最后，再经过一步组合操作，将C中的所有上下文相关音素，转换成对应的三个状态的隐马尔科夫模型，此时新得到的加权有限状态接收器，就是最终的解码图D。整个过程用公式可以表示为

$$D = G \quad o \quad T_{Words \rightarrow CI\_phones} \quad o \quad T_{CI\_phones \rightarrow CD\_phones} \quad o \quad T_{CD\_phones \rightarrow HMMs} \quad (2-20)$$

得到整个解码图之后，再通过维特比算法，找到最优路径，得到最终的识别结果。

### 2.1.5 评价指标

对于语音识别任务来说，比较常用评价指标为词级别错误率(Word Error Rate)。对于语音识别的结果，总共包含三种类型的错误，分别是替换错误、删除错误、插入错误。从字面意思上来理解，替换错误指的是原本句子里的某个词在识别结果里被错误替换成其它的词；删除错误指的是原本句子里出现的某个词，在识别结果里被错误地去掉了；插入错误错误指的是识别结果里出现了原本句子里没有出现过的某个词。这三种错误无论出现哪一种，都会使得识别性能下降。所以语音识别中使用的各种算法的目的，就是要尽量降低这三种错误出现的可能，也就是要降低词级别错误率。词级别错误率的公式表述如下：

$$WER = 100 \times \frac{count(Deletion) + count(Insertion) + count(Substitution)}{count(Words)} \quad (2-21)$$

其中， $count(words)$ 代表原始句子的单词总数， $count(Deletion)$ 、 $count(Insertion)$ 、 $count(Substitution)$ 分别代表识别结果中出现删除错误、插入错误、替换错误的个数。

## 2.2 Kaldi工具箱

Kaldi是一款基于Apache License v2.0开放协议的开源语音识别工具箱。相比于其它现存的语音识别工具箱，Kaldi工具箱的特点可以总结为以下几个方面：

1. 整合了有限状态转换器(Finite State Transducers)，Kaldi工具箱引入了外部库OpenFst<sup>[9]</sup>，OpenFst是一款开源的处理加权有限状态机的库函数。
2. 支持矩阵运算操作。Kaldi基于标准的线性代数库BLAS和LAPACK，构建了高效的矩阵算法库。

3. 算法在设计上注重通用性和可扩展性。
4. 基于Apache v2.0协议，更加开放。
5. 提供了很多语音识别标准数据库的完整解决方案。比如LDC机构提供的WSJ(Wall Street Journal)数据库、RM(Resource Management)数据库等都有完整的代码可以从头搭建语音识别系统。

图 2-8 展示了Kaldi工具箱的整体架构。Kaldi库函数里面的所有模块可以被分成两大部分，其中一部分使用外部库OpenFst；另外一部分使用线性代数运算库函数BLAS/LAPACK。只有一个模块'DecodableInterface'同时使用了这两个外部库函数。Kaldi库函数的所有功能都是通过用C++实现的命令行工具调用的，每个工具都可以通过指定特定参数实现不同的功能，并且工具都被设计可以通过管道来读取或写入标准的输入输出流，便于组合不同的工具实现复杂的功能。

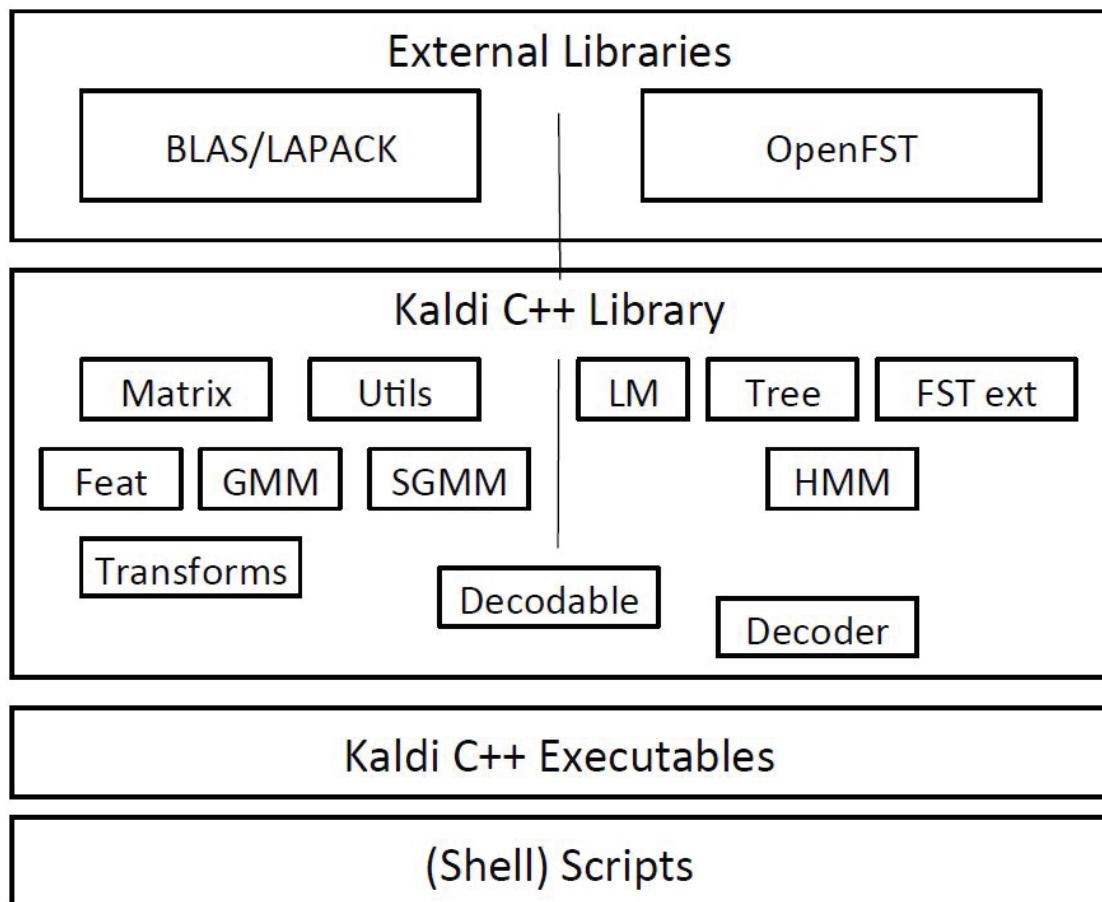


图 2-8 Kaldi 工具箱各个组成部分框图

从功能的角度来看，Kaldi工具箱可实现以下语音识别的相关操作：

- 声学特征提取。Kaldi目前已支持多种类型的声学特征，包括MFCC、PLP、filter-bank、bottle-neck features等。同时也提供了多种线性变换和仿射变换，例如：VTLN(Vocal Tract Length Normalization)、CMVN(Cepstral Mean and Variance Normalization)、LDA(Linear Discriminant Analysis)、HLDA(Heteroscedastic Linear Discriminant Analysis)等。
- 声学模型训练。Kaldi工具箱主要关注于声学模型的训练，目前已实现了包括GMM、SGMM、DNN、RNN、LSTM在内的几乎所有声学模型，同时也提供了很多判别式训练方法，比如MPE(Minimum Phone Error)、MMI(Maximum Mutual Information)等，另外还包括VTLN(Vocal Tract Length Normalization)、fMLLR(feature Maximum Likelihood Linear Regression)等在内的说话人自适应算法。
- 语言模型训练。Kaldi本身并不提供语言模型的训练工具，语言模型的训练可以使用IRSTLM或者SRILM等开源工具实现。由于Kaldi工具箱使用基于FST(Finite State Transducer)框架的训练和解码方法，一般常用的语言模型以ARPA格式存储，因此Kaldi提供了将ARPA格式的数据转成FSTs的相应工具。
- 构建解码图。Kaldi构建解码图的过程基于WFST<sup>[8]</sup>，Kaldi提供了一系列对解码图的优化操作，如Determinization、Minimization等。

本论文中所有声学模型的训练全部在Kaldi工具箱上完成。

## 第三章 声学模型训练

在2.1.2中，已经介绍了声学模型在语音识别中的作用，以及如何使用声学模型计算似然概率 $P(O|\Omega)$ 。本章主要介绍声学模型参数估计的问题，包括传统的GMM-HMM模型和目前广泛使用的DNN-HMM模型。

### 3.1 传统的GMM-HMM方法

在DNN-HMM模型兴起之前，GMM-HMM作为通用算法，一直在声学模型上占主导地位，几乎所有大型的实用语音识别系统都采用GMM-HMM作为其声学模型。GMM-HMM模型使用HMM模型对语音信号的时序特性建模，使用GMM模型计算每个隐马尔科夫模型的状态发射概率。本节首先介绍GMM模型的参数估计，接着再介绍GMM-HMM模型的训练过程。

#### 3.1.1 GMM模型的定义

GMM模型指的是混合高斯模型(Gaussian Mixture Model)。在介绍混合高斯模型之前，首先要介绍一下高斯分布。当一个连续随机变量 $x$ 的概率密度函数满足：

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-u)^2}{2\sigma^2}} \quad (3-1)$$

时，随机变量 $x$ 服从均值为 $u$ ，方差为 $\sigma^2$ 的高斯分布，记作 $x \sim \mathcal{N}(\mu, \sigma)$ 。因为 $x$ 为标量，因此又叫单变量的高斯分布。当 $x$ 变为向量 $X = (x_1, x_2, \dots, x_k)$ 时， $X$ 的高斯分布成为多变量的高斯分布，此时 $X$ 的概率密度函数满足

$$p(X) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)} \quad (3-2)$$

其中 $X$ 和 $\mu$ 均为 $k$ 维的向量， $\Sigma$ 为 $k \times k$ 的协方差矩阵。记作 $X \sim \mathcal{N}(\mu, \Sigma)$ 。

以上是单个的多变量高斯模型，接下来是混合的多变量高斯模型。

$$p(X) = \sum_{m=1}^M \frac{c_m}{(2\pi)^{\frac{k}{2}} |\Sigma_m|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_m)^T \Sigma_m^{-1} (X-\mu_m)} \quad (3-3)$$

其中， $M$ 表示混合的高斯模型的个数， $m$ 表示第 $m$ 个高斯模型， $\mu_m$ 表示第 $m$ 个高斯模型的均值向量， $\Sigma_m$ 表示第 $m$ 个高斯模型对应的协方差矩阵。当 $M$ 足够多时，混合的多变量高斯模型可以拟合任意分布。在语音识别中，可以使用混合高斯模型去拟合39维的MFCC向量。

### 3.1.2 GMM模型的参数估计

GMM模型的参数包括混合高斯数 $M$ 、每个高斯所占权重 $c_m$ 、每个高斯对应的均值向量 $\mu_m$ 和协方差矩阵 $\Sigma_m$ 。当随机变量 $X$ 的两个维度不相关时，其协方差为零。所以如果随机变量 $X$ 各维度之间都不相关，此时协方差矩阵为对角阵，可以大大减少模型参数。因此，语音识别在提取MFCC特征时，需要做DCT变换将频谱变为倒谱，可以降低MFCC各维度之间的相关性，保证后续的GMM可以使用对角阵。

EM算法(Expectation-Maximization)是训练GMM模型的主要工具。EM算法为迭代算法，每次迭代都包含两步：E步和M步。其核心思想是：对于有 $M$ 个成分的混合高斯模型，假设观测到的样本点 $X$ 由 $M$ 个成分中某一个成分 $GMM_m$ 的，但具体是哪一个成分并不能观测到。如果能够得到 $x$ 具体是由哪个成分得到的，那么整个模型就不再包含隐变量，可以直接使用最大似然估计(MLE)得到均值向量 $\mu_m$ 以及方差矩阵 $\Sigma_m$ 。EM算法不直接指定 $x$ 具体由哪个成分得到，而是在算法的E步时，计算每个高斯成分对这个样本点的占有概率。得到这些后验概率之后，整个模型不再包含隐变量，接下来就是M步。通过最大似然估计，更新参数 $c_m$ 、 $\mu_m$ 、 $\Sigma_m$ 。EM算法的具体公式如下。

E步：

$$h_m^{(j)}(t) = \frac{c_m^{(j)} \mathcal{N}(X^{(t)}; \mu_m^{(j)}, \Sigma_m^{(j)})}{\sum_{i=1}^M c_i^{(j)} \mathcal{N}(X^{(t)}; \mu_i^{(j)}, \Sigma_i^{(j)})} \quad (3-4)$$

M步：

$$c_m^{j+1} = \frac{1}{N} \sum_{t=1}^N h_m^j(t) \quad (3-5)$$

$$\mu_m^{j+1} = \frac{\sum_{t=1}^N N h_m(j)(t) X^{(t)}}{\sum_{t=1}^N h_m^{(j)}(t)} \quad (3-6)$$

$$\Sigma_m^{(j+1)} = \frac{\sum_{t=1}^N h_m^{(j)}(t) [X^{(t)} - \mu_m^{(j)}] [X^{(t)} - \mu_m^{(j)}]^T}{\sum_{t=1}^N h_m^{(j)}(t)} \quad (3-7)$$

其中， $M$ 表示共有 $M$ 个高斯成分、 $N$ 表示样本点个数、 $X^{(t)}$ 表示第 $t$ 个样本点、 $m$ 表示第 $m$ 个高斯成分、 $c_m$ 为第 $m$ 个高斯所占权重、 $\mu_m$ 为第 $m$ 个高斯的均值向量、

$\Sigma_m$ 为第m个高斯的协方差矩阵。

由于第一次迭代时，需要已经各个高斯成分的参数，因此使用EM算法时需要初始化，并且，EM算法收敛于局部最优。

### 3.1.3 GMM-HMM模型训练

GMM-HMM模型训练使用前后向算法，又称作Baum-Welch算法。Baum-Welch算法是EM算法在训练HMM模型上的具体应用。用前后向算法估计HMM模型状态转移概率的整体流程如下：

1. 首先通过前向算法，计算 $\alpha(S, t)$ 和 $P(X)$ 。
2. 再通过后向算法，计算 $\beta(S, t)$ 。
3. 对于每一个弧 $S \rightarrow^{x_t} S'$ 及时间步 $t$ ，计算这个弧在时间步 $t$ 产生的观测 $X = X_t$ 的概率

$$c(S \rightarrow^{x_t} S', t) = \frac{1}{P(X)} \times P_{S \rightarrow^{x_t} S'} \times \alpha(S, t - 1) \times \beta(S', t) \quad (3-8)$$

4. 对所有时间步下由状态 $S$ 跳转到 $S'$ 的所有情况求和。

$$c(S \rightarrow S') = \sum_{t=1}^T c(S \rightarrow^{x_t} S', t) \quad (3-9)$$

5. 最后，根据最大似然估计，从状态 $S$ 到状态 $S'$ 的状态转移

$$P_{S \rightarrow S'}^{MLE} = \frac{c(S \rightarrow^{x_t} S')}{\sum_{X, S'} c(S \rightarrow^{x_t} S')} \quad (3-10)$$

以上为使用Baum-Welch算法计算HMM状态转移概率的过程。

尽管GMM-HMM模型被成功应用到声学建模中，但GMM-HMM框架本身依然存在很多问题，比如HMM模型本身的条件独立假设的限制。随着神经网络模型理论及深度学习技术的发展繁荣，目前DNN模型已经取代了GMM模型，并且人们也逐渐使用RNN模型(循环神经网络)替换HMM模型，来对语音信号的序列特性建模。

## 3.2 目前广泛使用的DNN-HMM方法

在DNN模型被广泛应用到语音识别之前，GMM-HMM一直在声学模型上占主导地位。2006年之后，Hinton等人提出了DNN预训练方法，通过预训练受限玻尔兹曼机来初始化DNN模型，使得训练深度的神经网络成为可能。之后，

DNN模型逐步取代GMM模型，DNN-HMM成为声学建模的主流方法。与GMM-HMM相比，DNN-HMM的优势在于使用DNN这一强大的判别模型替代传统的GMM模型来计算HMM的状态发射概率。图 3-1 为DNN-HMM声学模型的结框图。

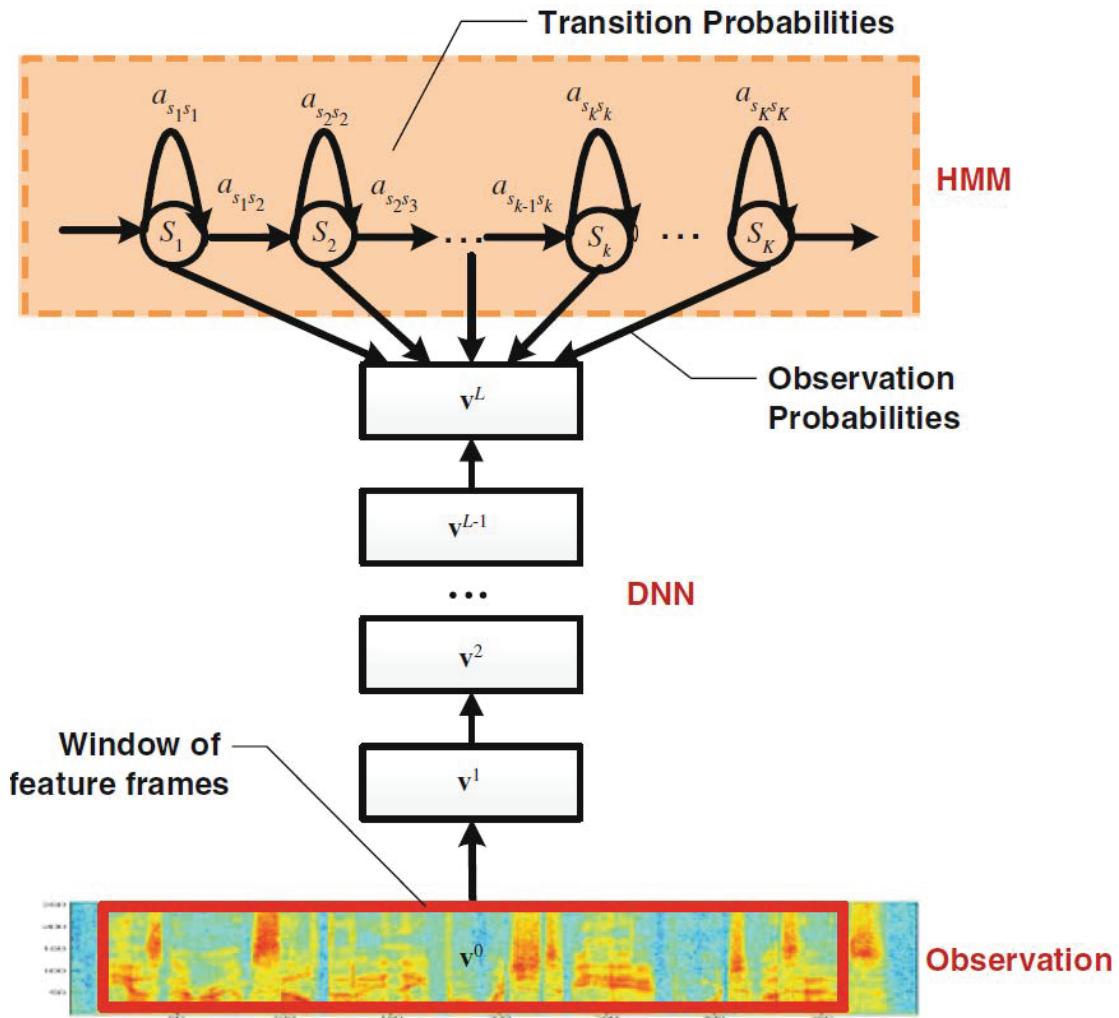


图 3-1 DNN-HMM声学模型结构图<sup>[10]</sup>

### 3.2.1 DNN-HMM模型训练步骤

DNN-HMM模型的训练可以使用内嵌的维特比算法(Embedded Viterbi)实现。DNN-HMM模型总共包含三个部分：DNN模型、HMM模型、状态的先验概率分布。由于DNN-HMM与GMM-HMM使用同样的三音素模型和HMM模型，所以在训练DNN-HMM之前，首先要训练GMM-HMM模型。由于DNN-HMM在模型训练时需要的每个句子的音素标注是由GMM-HMM模型对每个句子做强制对

齐时得到的，因此得到一个好的GMM-HMM模型对训练DNN-HMM模型十分重要。训练好GMM-HMM模型之后，使用GMM-HMM模型对训练集的每个句子使用维特比算法得到状态对齐信息。再由状态对齐信息通过映射关系得到每个状态对应的senone。再通过切分信息将senone映射到实际的输入特征上。同时通过senone到特征的映射得到每个状态的先验概率分布。复制已经得到的GMM-HMM模型中的HMM模型，用作DNN-HMM模型中的HMM。接下来是DNN模型的预训练，通过训练深度置信网络(DBN)，将DBN的模型参数作为DNN的初始化参数，接着使用反向传播(Back Propagation)算法进行随机梯度下降，最终得到DNN模型。至此，DNN-HMM模型参数已全部得到。

## 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已经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。<sup>[11]</sup>是较早的使用DTW做拉萨方言孤立词识别的，<sup>[12][13]</sup>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。<sup>[14][15]</sup>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<sup>[14][15]</sup>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句拉萨口语音素平衡句，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。图表 4-1 是发音字典所用声韵母集合。

## 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。在搭建基准系统之前，我们尝试了两种不同的DNN声学模型框架，一种是DNN-HMM<sup>[16]</sup>；另外一种是Tandem方法<sup>[17]</sup>。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。

Lhasa Tibetan initials							
p	c	ts	tɕ	ɳ	s	h	
ph	ch	tsh	tɕh	ɳj	tʃ	x	
t	k	tʂ	m	l	ʂ	w	
th	kh	tʂh	n	f	ç	j	
ç							
Lhasa Tibetan finals							
i	y	o:	io	an	oŋ	op	u?
e	ø	u:	im	on	uŋ	up	ir
a	ɛ	y:	em	yn	ip	i?	er
ə	i:	ø:	am	iŋ	ep	e?	ar
o	e:	ɛ:	om	en	ap	a?	or
u	a:	iu	um	aŋ	əp	o?	ur

图 4-1 拉萨发音字典声韵母集合

### 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。训练过程中，特征空间最大似然线性回归(fMLLR)方法用来做说话人归一化，消除不同说话人对识别结果的影响。训练阶段，90%的训练数据用来作为模型训练数据，10%的训练数据作为验证集数据，用来选择学习速率等模型参数。整个的CD-DNN-HMM模型训练过程如图4-2所示。

### 4.3.2 Tandem

Tandem方法是将深度神经网络模型与传统的GMM-HMM模型融合的一种方式，在DNN-HMM模型中，DNN的隐层用做非线性特征变换，输出层(通常为softmax层)用做分类器。而在Tandem方法中，DNN模型只用做特征提取，最后的声学模型还是传统的GMM-HMM。而DNN提取的特征并不直接使用输出层的输出，而是用一个节点数明显少于其它隐层的中间隐层的激励函数的输出作为

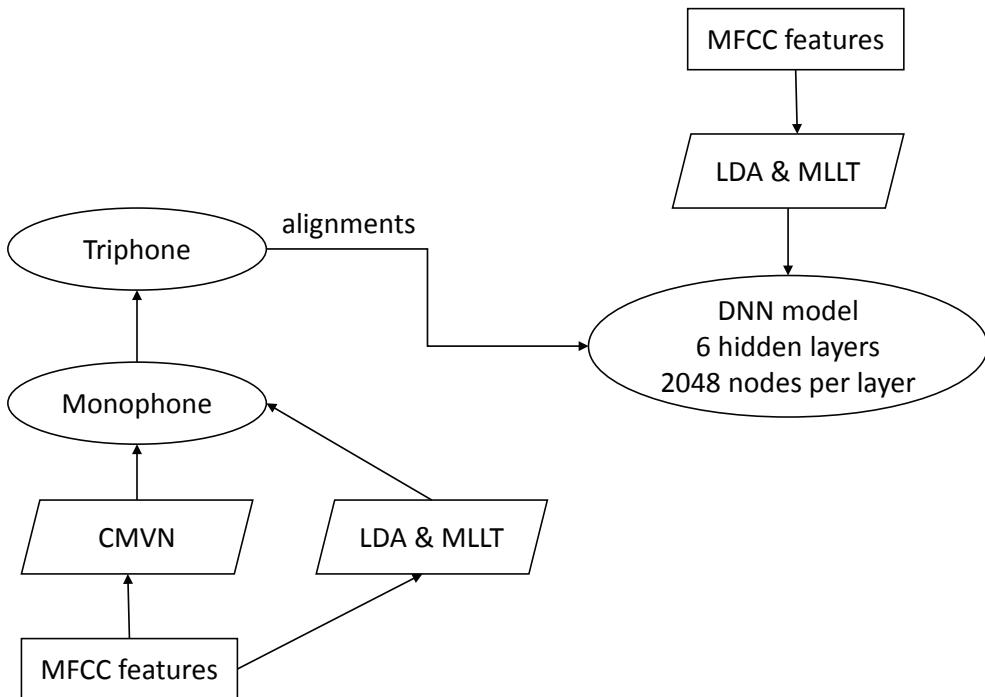


图 4-2 CD-DNN-HMM训练过程

输出特征。使用中间隐层输出的好处是，中间隐层的节点个数独立于输出层节点个数，可以随意设置，这样方便控制特征的维度。一般情况下，DNN的bottle-neck层的输出会拼接到MFCC或者PLP特征上，并且使用PCA或者HLDA降维并去除相关性，之后再训练GMM-HMM。

在本论文中，我们使用一个带有bottle-neck层的前馈神经网络，总共有三个隐层，中间的那个隐层作为bottle-neck层。bottle-neck层的节点数为42，其余隐层节点数为1024。隐层节点的激励函数使用双曲正切函数。整个Tandem方法的详细过程如图 4-3 所示。

### 4.3.3 两种建模方法的音素级识别结果

为了单纯地比较两种声学建模方法的准确率，而不用过多考虑语言模型的影响，我们首先使用音素级别错误率作为评价指标。此时用到的语言模型的训练数据全部来自于训练数据的音素标注。下表 4-1 是两种声学建模方法的音素级别的错误率。

为了更直观地表示两种声学建模方法的识别错误类型，本文在这里对两种方法分别画出了音素混淆矩阵。 $x$ 轴代表音素的标签， $y$ 轴代表音素的识别结果。颜色越亮代表音素 $x$ 被识别成 $y$ 的概率越大。对角线上的每个点表示音素被正确

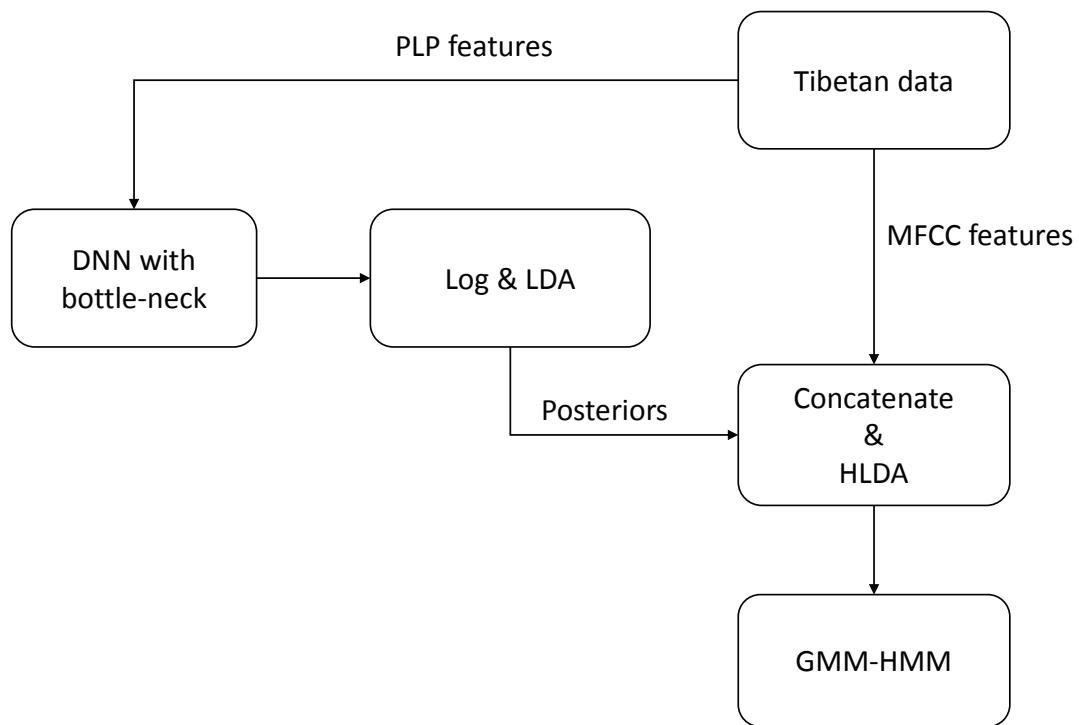


图 4-3 Tandem方法实现过程

表 4-1 两种声学建模方法对应音素级别错误率

Acoustic Model	PER
CD-DNN-HMM	20.84%
Tandem Approach	21.48%

识别的概率。图 4-4 和图 4-5 分别是 CD-DNN-HMM 和 Tandem 对应的混淆矩阵：

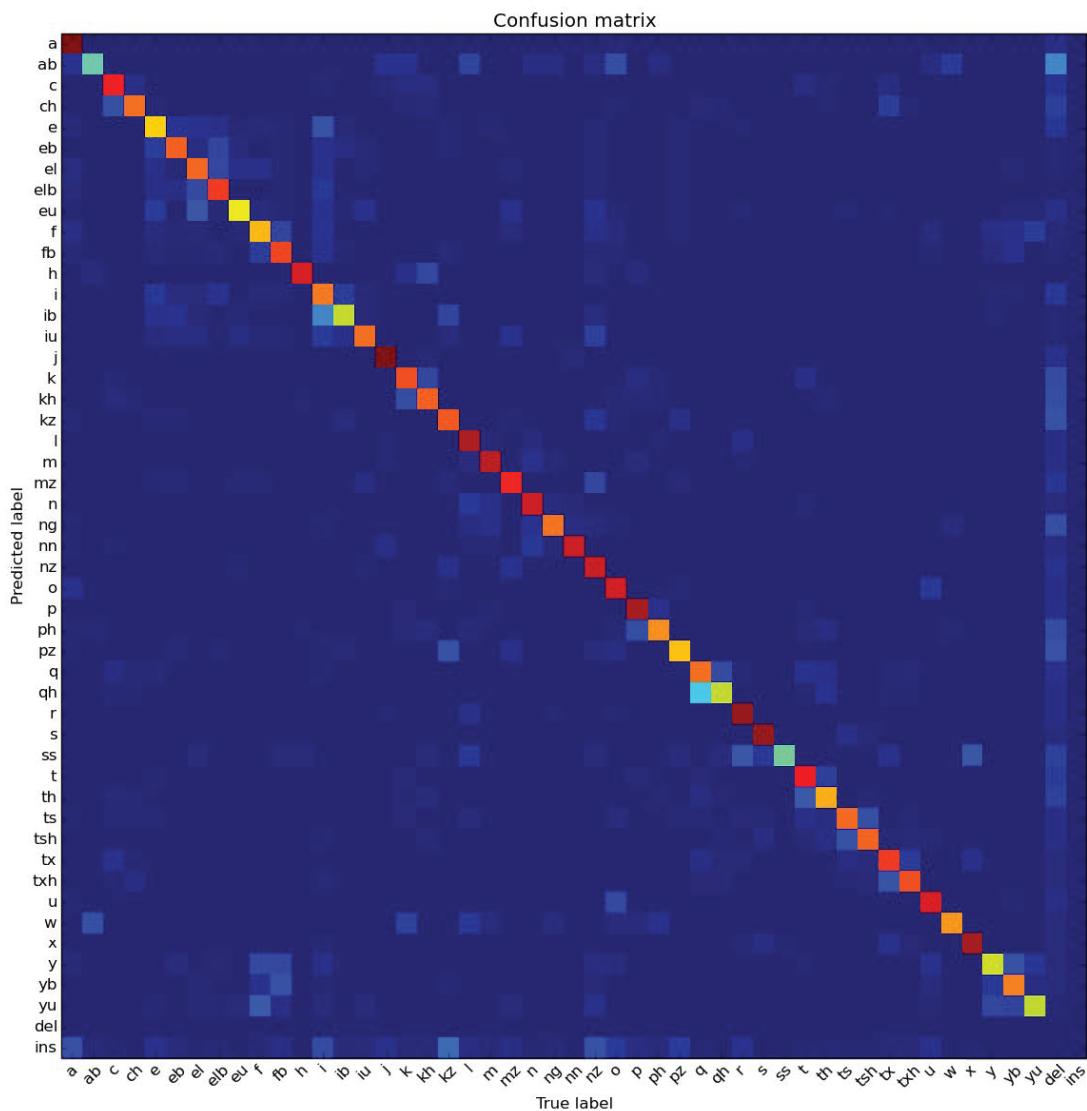


图 4-4 CD-DNN-HMM 音素识别结果混淆矩阵

通过比较两个混淆矩阵，我们可以看到两种方法识别结果的错误类型大体一致，在DNN-HMM中出现的比较严重的错误类型，在Tandem方法中也同样出现了。根据以上的识别结果和错误类型分析，我们决定选用结果稍好的CD-DNN-HMM方法作为基准系统的声学模型训练方法。

#### 4.3.4 CD-DNN-HMM基准系统

选定了CD-DNN-HMM作为声学模型之后，为了搭建实用的语音识别系统，还必须使用有效的语言模型。因为我们目前没有较好的藏文分词工具，所以系

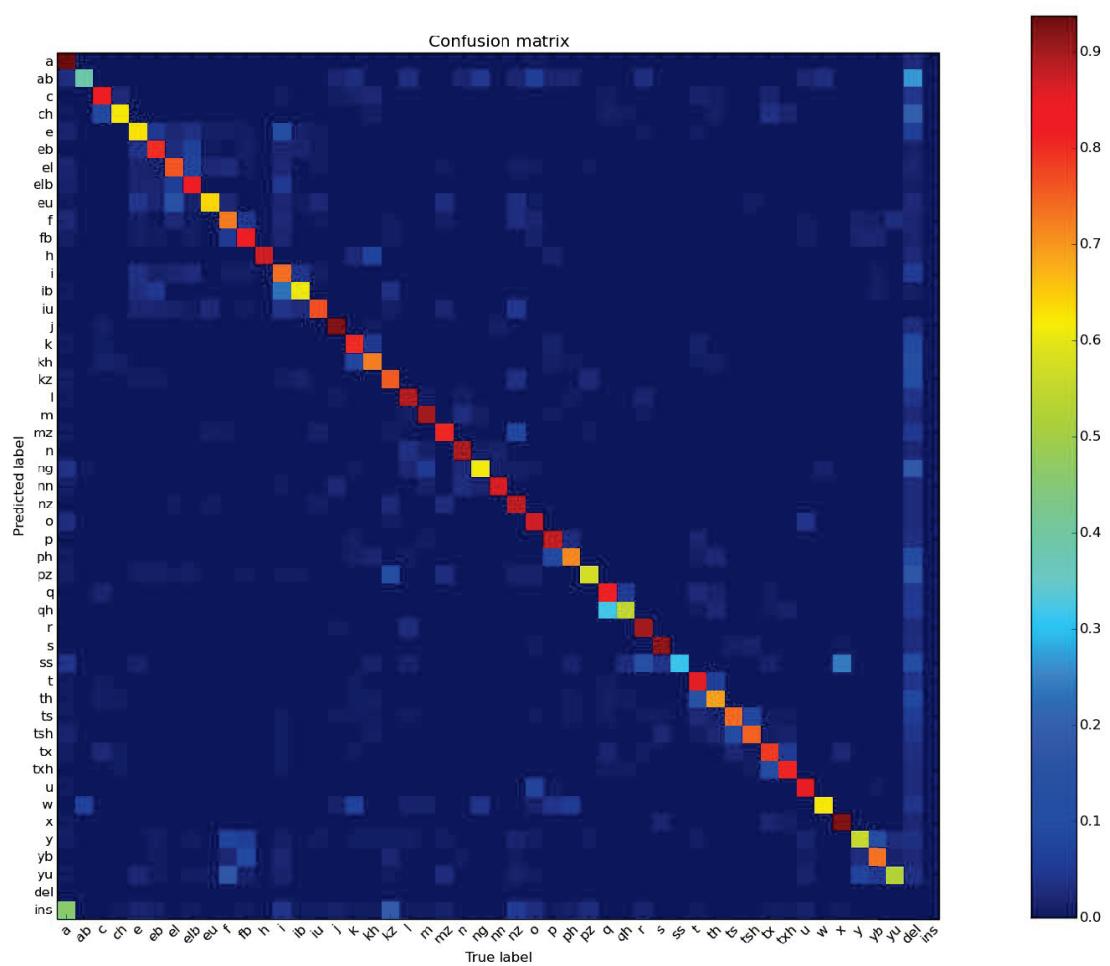


图 4-5 Tandem 音素识别结果混淆矩阵

统使用字级别的语言模型并采用字级别的错误率作为评价标准。由于目前训练声学模型使用的数据为拉萨口语句子，很难搜集和音频数据主题相关的口语文本语料。目前使用的语言模型的训练数据包含两个部分，一部分是从维基百科上爬取的藏语文本数据；另外一部分是藏族五省区中学课本的一部分。总共包括14,430个藏语句子。语言模型使用的是三元语法模型，平滑方法选用的是Kneser-Ney方法。表 4-2 是基准系统的字级别错误率：

表 4-2 基准系统字级别错误率

Acoustic Model	CER	Details
CD-DNN-HMM	33.97%	8,190/24,108 [246 ins; 154 del; 7,790 sub]

通过检查识别结果，我们发现有很多由于同音字造成的替换错误。图 4-6 是测试集中随机挑选的一个句子及其对应的识别结果，其中两处替换错误都被红色标识出来了。对于第一处替换错误，原文字和替换文字是一对儿同音字，组成两个字发音的声韵母在发音字典里是相同的，因此导致了识别错误。同音字替换错误本来是可以通过语言模型解决的，但是由于目前使用的语言模型的训练数据缺少主题相关的口语语料，导致语言模型的精度较差。这也是目前拉萨方言语音识别普遍存在的问题。然而，图 4-6 中的同音字替换错误的例子，两个同音字的声调其实是不同的，所以我们可以利用拉萨方言的声调信息来提高识别的准确度。

TF207\_0002 རྒྱ ། ག རྩ ག རྩ

TF207\_0002 རྒྱ ། ག རྩ ག རྩ

图 4-6 标号为TF207\_0002的句子，上一行为文本，下一行为识别结果

#### 4.4 拉萨方言声调系统

在利用声调信息之前，我们首先需要了解清楚拉萨方言的声调。和普通话的声调类似，拉萨方言的声调也是具有词汇意义的，即声调可以用来区分同音

字。为了能够对声调建模，我们需要知道拉萨方言的声调总共有多少种类型，每种类型的声调曲线形状是什么样的。

#### 4.4.1 拉萨方言的四个声调类型

藏语属于汉藏语系，缅藏语组。藏语有三大方言，分别是：拉萨方言、康巴方言和安多方言。其中康巴方言和拉萨方言被认为是有调语言，而安多方言被认为是无调语言。在这三大方言中，拉萨方言的声调特性更加丰富。然而，有关拉萨方言究竟该划分为多少个声调还存在争议。其中，已经被广泛认可的是拉萨方言中存在高低调的区别。根据<sup>[18]</sup>所述，拉萨方言的声调韵律在很大程度上由音节的类型决定。从音韵学角度来说，拉萨方言可以被认为共有两个、四个、六个，甚至八个声调。声调个数的不确定性给利用声调信息带来了困难。在本论文中，我们的目的是找到一个能够有效利用声调信息提高识别准确率的方法。因此，我们基于<sup>[19]</sup>设计了四个声调的拉萨方言声调系统。根据音韵学原理，我们使用五度值的方法<sup>[20]</sup>来表示不同的声调类型。四个声调可分别表示为：55(高平调)、13(升调)、51(降调)、132(升降调)。由于拉萨方言的韵母是携带声调的主要部分，因此，我们将音素表中的所有韵母扩充成四个不同声调的韵母，声母保持不变。最后得到的音素集合总共包含29个声母和192个带调韵母。图4-7和图4-8展示的是从训练集中随机挑选的一个句子，图4-7是女性发音人录制的，图4-8是男性发音人录制的。每个图中均包含声音波形、语谱图、音素标注。语谱图中的细实线表示的是声调曲线，音素标注使用的是带调的音素集合。

### 4.5 声调特征提取

因为我们是通过声调曲线轮廓去区分不同声调的，所以对声调建模实际上是pitch-tracking问题。pitch并不是纯粹客观的物理性质，它是声音的主观心理声学属性。而基频F0(Fundamental frequency)作为周期性信号的固有属性，与人对声音的pitch感知联系十分紧密。因此，几乎所有的pitch-tracking算法都是去估计声音的基频值<sup>[21]</sup>。在本论文中，我们总共使用了两种提取基频值的方法，一个是SAC方法<sup>[22]</sup>，另外一个是Kaldi-pitch方法<sup>[23]</sup>。这两个方法在pitch-tracking标准测试集Keele数据集<sup>[24]</sup>上都表现出了很好的性能。

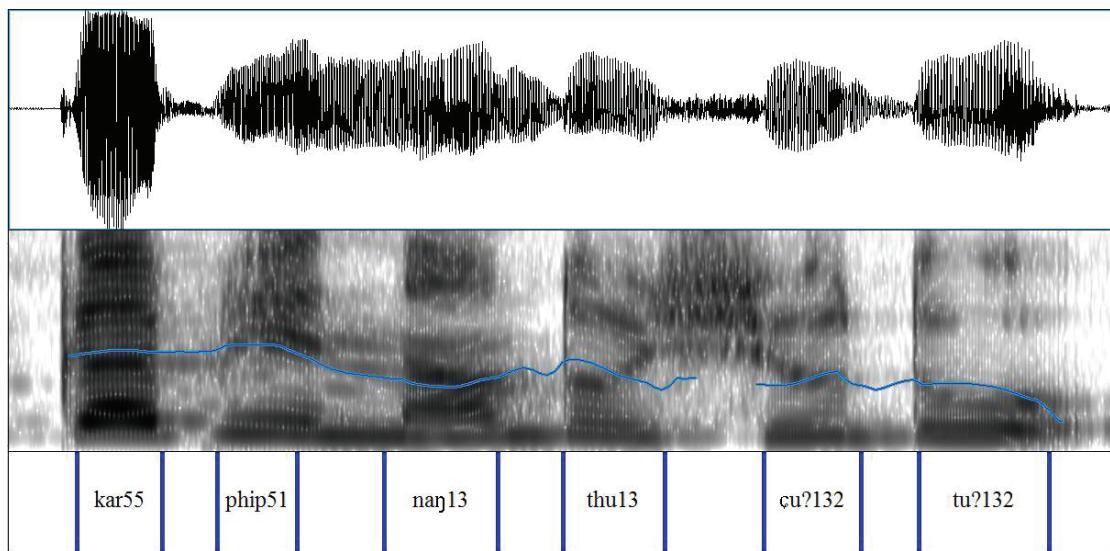


图 4-7 女性发音人基频曲线及音素标注

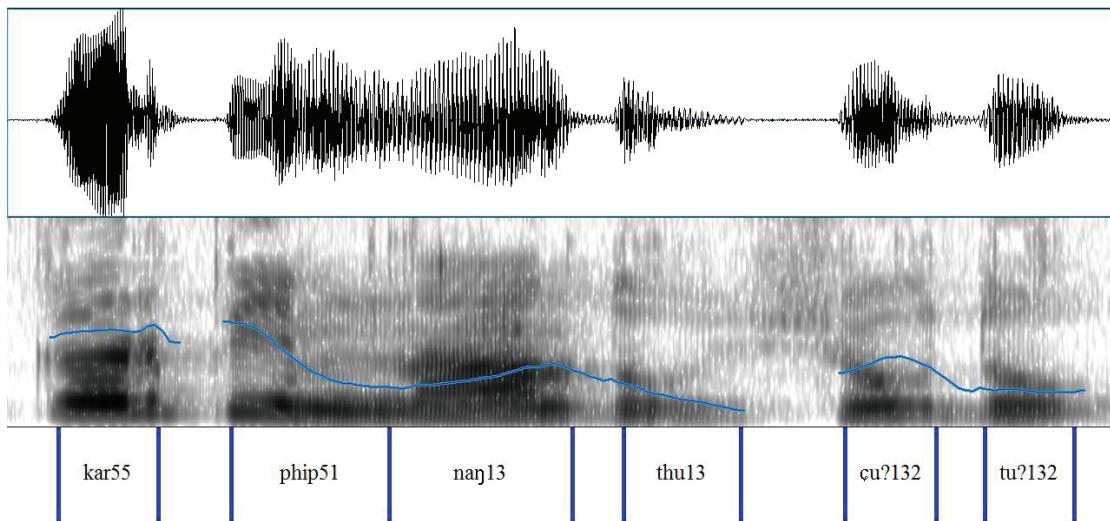


图 4-8 男性发音人基频曲线及音素标注

### 4.5.1 SAcC方法

SAcC全称是Subband Autocorrelation Classification的缩写，该方法首先将音频信号 $a[n]$ 通过一个带有48个带通滤波器的滤波器组，得到48个子带信号 $b_k[n], k = 1, 2, \dots, 48$ 。接着计算每个子带信号归一化的自相关系数(Normalized Autocorrelation)，然后使用PCA对自相关系数降维，最后用降维后的自相关系数训练多层感知机(MLP)模型。MLP的输出可以用来估计给定观测下，pitch为某一值的后验概率。最后，基于MLP的输出，使用Viterbi算法搜索每一帧最佳的基频值，同时判断该帧是否为静音帧。详细的SAcC方法描述可以参考<sup>[22]</sup>。

### 4.5.2 Kaldi-Pitch方法

Kaldi-Pitch方法<sup>[23]</sup>属于数据驱动方法，是对经典的基频提取方法Getf0方法<sup>[21]</sup>的改进版本。Getf0方法通过计算归一化的互相关函数得到音频信号每一帧的一系列基频备选值，之后使用DP(Dynamic Programming)算法，根据局部和全局的信息，得到每一帧的最佳基频值，对于静音帧，直接将基频值赋成0。Kaldi-Pitch方法与Getf0方法的最大区别在于，Kaldi-Pitch并不直接决定每一帧是否为静音帧。对于静音帧，Kaldi-Pitch方法会将其基频赋成一个非零值，目的是保持基频曲线的连续性。但同时，Kaldi-Pitch方法引入了发声概率这个变量，用来描述该帧是非静音帧的概率。<sup>[23]</sup>的实验表明，用连续的基频曲线加上每一帧的发声概率作为声调信息，结合MFCC特征，在BABEL这个标准数据集上对多个语种的语音识别都取得了不错的识别效果。和Getf0方法一样，Kaldi-Pitch用来计算基频备选值的方法同样是基于归一化的互相关函数。假设 $x_p, p = 0, 1, 2, \dots$ 是采样后的语音信号，采样周期 $T = 1/F_s$ ，时间窗长度为 $\omega$ ，窗移长度为 $t$ ，每个时间窗内的采样点个数 $n = \omega/T$ ，窗与窗之间相距的采样点个数 $z = t/T$ 。那么，对于第*i*帧信号，延迟系数为*k*时，其对应的互相关系数 $\phi_{ik}$ 的计算公式如下：

$$\phi_{ik} = \frac{\sum_{j=m}^{m+n-1} x_j x_{j+k}}{\sqrt{e_m e_{m+k}}} \quad (4-1)$$

$$e_j = \sum_{i=j}^{j+n-1} x_i^2 \quad (4-2)$$

其中， $m = i * z$ ，i代表帧的标号，k代表延迟系数。

对于每一帧信号*i*，NCCF的局部极大值点就是基频值的备选。根据上述公式，我们可以知道 $-1.0 \leq \phi_{ik} \leq 1.0$ ，并且当延迟系数*k*接近基频的正整数倍时，

$\phi_{ik}$ 接近于1.0。

假设c是给定某一帧的NCCF值，其绝对值 $a = abs(c)$ 。以下是通过分析Keele数据库的 $\log(\frac{count(voiced)}{count(unvoiced)})$ 曲线得到的公式：

$$h = -5.2 + 5.4e^{7.5(a-1)} + 4.8a - 2e^{10a} + 4.2e^{20(a-1)} \quad (4-3)$$

此处， $h$ 近似等于 $\log(\frac{count(voiced)}{count(unvoiced)})$ ， $p = \frac{1}{1+e^{-h}}$ 近似等于当前帧是非静音帧的概率。

### 4.5.3 声调相关特征

在本论文中，提取基频相关特征使用两种不同的pitch-tracking方法，分别是SAcC方法和Kaldi-Pitch方法。由SAcC方法计算的基频值只在非静音段不为零，造成基频曲线不连续，我们可以选择是否做插值平滑处理。所以，总共会有三种不同的基频相关特征，分别是Kaldi-F1、Interpolate-SAcC-F2、Raw-SAcC-F3。以下是对这三种特征的详细介绍。

尽管基频只在音频的有声段才有定义，但由于Kaldi-Pitch方法并不会直接将无声段的基频值赋为0，从而保证了基频曲线的连续性，因此不需要做额外的插值处理。Kaldi-Pitch方法的输出包括基频的对数值还有每一帧为非静音帧的概率。为了提取基频曲线的动态特征，我们使用当前帧的前两帧和后两帧计算了基频的delta值。因此，对于每一帧音频数据，总共有三维的基频相关特征，直接将这三维特征拼接到MFCC特征后面，得到Kaldi-F1特征。

对于SAcC方法，由于它直接将无声段的基频值赋成0，为了避免特征的不连续导致识别性能下降，此处对基频曲线做了插值和平滑处理，具体过程类似于<sup>[25]</sup>。选用的插值方法是piecewise cubic Hermite interpolating polynomial (PCHIP)<sup>[26]</sup>。选用的平滑方法是5-point moving average。与Kaldi-Pitch方法类型，SAcC方法同样提供每一帧为非静音帧的概率。因此，使用SAcC方法，对于每一帧，同样也是三维的基频相关特征，包括基频的对数值、非静音帧的概率、基频的delta值。将这三维的特征拼接到MFCC特征后面，得到Interpolate-SAcC-F2特征。为了比较使用连续的基频曲线和不连续的基频曲线对声学建模的效果，此处还设置了Raw-SAcC-F3特征，同样也是将基频特征直接拼接到MFCC特征后面，但此处的基频相关特征只包括未处理的基频值F0、非静音帧的概率。

## 4.6 加入声调信息的识别系统

本论文中，引入声调信息主要通过两种方式，一种是将音素集合中的韵母扩充成带调韵母的形式；另外一种就是使用声调相关特征训练声学模型。为了评价带调音素集合和声调相关特征的有效性，此处设置了四组对比试验。首先是使用带调的音素集合加上MFCC特征训练声学模型；接下来将音素集合变为带调音素集合，将MFCC分别替换为Kaldi-F1、Interpolate-SAcC-F2以及Raw-SAcC-F3。声学模型和语言模型的训练方法和参数配置与4.3.4中的基准系统相同。表 4-3 为四个不同的声学模型以及基准系统在测试集上的字级别错误率。

表 4-3 四个不同声学模型及基准系统字级别错误率

Different Models	CER
Non-tonal Phone + MFCC	33.97%
Tonal Phone + MFCC	31.91%
Tonal Phone + Kaldi-F1	30.20%
Tonal Phone + Interpolate-SAcC-F2	31.30%
Tonal Phone + Raw-SAcC-F3	31.91%

从表中的识别结果可以看出，通过使用带调的音素集合，即使是特征层面不做任何改变，相比基准系统能获得相对6.1%的性能提升。当使用带调音素集合，同时将MFCC换成Kaldi-F1特征时，累积的性能提升为11.1%。当把Kaldi-F1特征换成Interpolate-SAcC-F2特征时，性能有所下降，但依然要好于带调音素集合加上MFCC特征的情况。<sup>[23]</sup>中展示了使用Kaldi-Pitch方法得到的特征相比于SAC方法要更适合用来训练SGMM模型，而本论文中的结果表明，对于DNN模型而言，Kaldi-Pitch方法依然要优于SAC方法。另外，当直接使用原始的基频值加上非静音帧概率，而不做插值和平滑处理时，识别结果和使用带调音素结合加上MFCC特征相同。这表明此处的特征层面的声调信息对识别没有起到帮助。从识别结果中可以看到，利用声调信息确实可以提高识别准确度，但是目前总体的识别准确度并不是很高。主要原因是目前训练语言模型的训练数据与测试集的口语句子不匹配，导致语言模型性能较差。

### 4.6.1 系统融合

除了尝试不同的声调特征来训练声学模型，本论文还尝试了使用基于Minimum Bayes Risk(MBR)的lattice combination<sup>[27]</sup>，此方法将不同系统模型对同一个句子得到的lattice融合为一个统一的lattice。此处总共做了四种不同的系统融合，表 4-4 为系统融合后得到的识别结果。

表 4-4 不同类型的系统融合字级别错误率

Combining Different Systems	CER
Kaldi-F1 + Interpolate-SAcC-F2	28.92%
Kaldi-F1 + Raw-SAcC-F3	28.97%
Interpolate-SAcC-F2 + Raw-SAcC-F3	29.87%
All three systems	28.54%

从上表中的结果可以看出，将三个不同的系统融合在一起能得到最高的识别性能，Kaldi-F1系统和Interpolate-SAcC-F2系统之间的融合要优于Kaldi-F1系统和Raw-SAcC-F3系统之间的融合。无论哪一种系统融合方式，其结果都要优于任何单独的系统。基于以上结果，可以推测，不同的系统模型之间可以优势互补，尤其是当两种模型使用不同的基频提取方法时，融合后的效果提升更加明显。

## 第五章 总结与展望

语音识别技术发展至今，在部分应用场景下已达到实用阶段。尤其是对于类似英语、汉语等已经被语音识别领域深入研究的语言，语音识别系统的性能已经可以和人耳媲美。但对于很多语料资源匮乏的语种来说，由于其训练数据获取困难，加上缺少语言的语音学和语言学的专业知识，对于这些语种的语音识别研究已远远落后于目前的流行技术。

拉萨方言作为少数民族语言，同样属于语料匮乏语言，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。与此同时，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，最终建立了拉萨方言带调音素集合。在特征层面，尝试了使用两种不同的基频提取方法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用来训练声学模型。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。实验的训练数据31.9小时，测试集2.41小时。识别结果表明，对于DNN模型，使用区分声调的音素集合加上MFCC特征，字级别错误率相对下降6.1%；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。

对于拉萨方言的声调类型，目前采用的是四个声调，但这是基于藏语言学家的先验知识得到的。后续可能通过分析声调的功能负载，去确定拉萨方言究竟该采用多少个声调比较合适。并且，目前的DNN-HMM声学模型已逐步被LSTM方法取代，所以后续的识别系统可以尝试使用基于LSTM的声学模型。

## 参考文献

- [1] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit [C]. In IEEE 2011 workshop on automatic speech recognition and understanding, 2011.
- [2] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE transactions on acoustics, speech, and signal processing, 1980, 28 (4): 357–366.
- [3] Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. the Journal of the Acoustical Society of America, 1990, 87 (4): 1738–1752.
- [4] Juang B-H, Rabiner L, Wilpon J. On the use of bandpass liftering in speech recognition [J]. IEEE Transactions on acoustics, speech, and signal processing, 1987, 35 (7): 947–954.
- [5] Stevens S S, Volkmann J. The relation of pitch to frequency: A revised scale [J]. The American Journal of Psychology, 1940, 53 (3): 329–353.
- [6] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition [J]. IEEE transactions on acoustics, speech, and signal processing, 1978, 26 (1): 43–49.
- [7] Goodman J T. A bit of progress in language modeling [J]. Computer Speech & Language, 2001, 15 (4): 403–434.
- [8] Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition [J]. Computer Speech & Language, 2002, 16 (1): 69–88.
- [9] Allauzen C, Riley M, Schalkwyk J, et al. OpenFst: A general and efficient weighted finite-state transducer library [C]. In International Conference on Implementation and Application of Automata, 2007: 11–23.
- [10] Yu D, Deng L. Automatic Speech Recognition [M]. Springer, 2012.
- [11] YAO X, SHAN G-r, YU H-z. Research on Tibetan Isolated-word Speech Recognition System [J] [J]. Journal of Northwest University for Nationalities (Natural Science), 2009, 1: 011.
- [12] LI G-y, Meng M. Research on Acoustic Model of Large-vocabulary Continuous Speech Recognition for Lhasa Tibetan [J]. Computer Engineering, 2012, 5: 189–191.
- [13] Li G Y, Yu H Z. Large-Vocabulary Continuous Speech Recognition of Lhasa Tibetan [C]. In Applied Mechanics and Materials, 2014: 802–806.
- [14] Wang H, Zhao Y, Xu Y, et al. Cross-language speech attribute detection and phone recognition for Tibetan using deep learning [C]. In Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on, 2014: 474–477.
- [15] Zhao Y, Zhou N, Zhang L, et al. Shared speech attribute augmentation for English-Tibetan cross-language phone recognition [C]. In 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015: 539–543.

- 
- [16] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20 (1): 30–42.
  - [17] Hermansky H, Ellis D P, Sharma S. Tandem connectionist feature extraction for conventional HMM systems [C]. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2000: 1635–1638.
  - [18] Hu F, Xiong Z. Lhasa tones [C]. In *Proceedings of the 5th international conference on speech prosody*, 2010.
  - [19] 于道泉. 藏汉拉萨口语词典. 1983.
  - [20] Chao Y. A system of tone letters [J]. *Le MaiVtre PhoneTtique*, 1980.
  - [21] Kleijn W B, Paliwal K K. A robust algorithm for pitch tracking [J]. *Speech Coding and Synthesis*, Elsevier, New York, 1995.
  - [22] Lee B S, Ellis D P. Noise robust pitch tracking by subband autocorrelation classification [C]. In *Interspeech*, 2012: 707–710.
  - [23] Ghahremani P, BabaAli B, Povey D, et al. A pitch extraction algorithm tuned for automatic speech recognition [C]. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014: 2494–2498.
  - [24] Plante F, Meyer G, Ainsworth W. A pitch extraction reference database [J]. *Children*, 1995, 8 (12): 30–50.
  - [25] Lei X, Siu M-H, Hwang M-Y, et al. Improved tone modeling for Mandarin broadcast news speech recognition. [C]. In *INTERSPEECH*, 2006.
  - [26] Fritsch F N, Carlson R E. Monotone piecewise cubic interpolation [J]. *SIAM Journal on Numerical Analysis*, 1980, 17 (2): 238–246.
  - [27] Xu H, Povey D, Mangu L, et al. Minimum bayes risk decoding and system combination based on a recursion for edit distance [J]. *Computer Speech & Language*, 2011, 25 (4): 802–828.

## 发表论文和参加科研情况说明

### (一) 发表的学术论文

[1] XXX, XXX. XXXXXXXXXX.

### (二) 参与的科研项目

[1] XXX, XXX. XX.

## 致 谢

本论文的工作是在我的学业指导老师党建武教授的指导下完成的，党建武教授严谨的治学态度和科学的工作方法以及对科学的研究的热忱深深地影响了我，使我逐渐对科研产生兴趣。在此由衷地感谢党老师对我学业上的指导和生活上的关心和帮助。

感谢李雪威副教授对我的学习和生活上的指导与关怀。

王龙标教授、本多清志教授、冯卉副教授、Bruce Denby教授、魏建国副研究员、王洪翠老师、刘志磊老师悉心指导我有关课题实验中的细节问题，帮助我在科研的道路上摸索前进，在此向各位老师表示衷心的谢意。

感谢陪伴了我三年的研究生舍友，尤其是来自云南藏族的鲁茸江才同学，不仅在科研上给予我无私的帮助，也在平时的做人做事方面时刻影响我，让我学到了很多在课堂上和实验室里学不到的东西。

研究生的生活离不开实验室同学的陪伴，感谢同届的小伙伴们在生活上的关心和照顾；感谢更太加博士对我的藏语知识方面的无私帮助；感谢申彤彤、赵涔汐、原梦、郭震等师弟师妹对我的科研及生活中的关乎与帮助。

感谢远在美国读研的于洋，你对新鲜事物的好奇心时刻影响着我，让我明白兴趣是最好的老师。

最后，衷心地感谢我的父母和哥哥，是你们的支持让我能够在与世无争的校园里安心学习科研。