深度学习初稿撰写要求：

1. 不能直接粘贴拷贝网上或者其他书籍内容，需按照自己的理解撰写。
2. 图表不能用现有图表，需重新制作，图用Vision绘图，表用Word制表。
3. 每章后需整理本章参考文献
4. 不局限于提纲内容，有修改或者新加内容随时和我讨论
5. 图文并茂，尽量通俗易懂。

**第四章 循环神经网络（Recurrent Neural Network，RNN）（景琨）**

4.1 循环神经网络

4.2 循环神经网络的训练方法

4.3 长短时记忆网络

4.4 双向循环神经网络

4.5循环神经网络的典型应用（简介）

**第七章 其他典型深度学习方法（景琨，王义宇，李航航）**

7.2 胶囊网络（景）

7.4 记忆网络（景））

7.5 增强深度学习（景）

**第八章 深度学习中的正则化（景琨，王义宇）**

8.1 正则化的概念（景）

8.2 L1和L2正则化（景）

8.5 Dropout和Dropconnect（景）

**第十一章 深度学习在语音识别中的典型应用（景琨、李航航）**

**需有数据集和代码**

11.2 语音识别（景）

11.3 声纹识别（景）

# 第四章 循环神经网络

## 4.1 循环神经网络

## 4.2 循环神经网络的训练方法

## 4.3 长短时记忆网络

## 4.4 双向循环神经网络

## 4.5 循环神经网络的典型应用（简介）

# 第七章 其他典型深度学习方法

## 7.2 胶囊网络

## 7.4 记忆网络

## 7.5 增强深度学习

# 第八章 深度学习中的正则化

## 8.1 正则化的概念

## 8.2 L1和L2正则化

## 8.5 Dropout和Dropconnect

# 第十一章 深度学习在语音识别中的典型应用

## 11.1 语音识别

语音识别（Automatic Speech Recognition，ASR）所要解决的问题是让计算机能“听懂”人类的语音，将语音转化成文本。语音识别是实现智能的人机交互的前沿阵地，是完全机器翻译，自然语言理解等的前提条件。语音识别是一门交叉学科，它所涉及的领域包括：信号处理、模式识别、概率论和信息论、发声机理和听觉机理、人工智能等等。

近年来，语音识别技术取得显著进步，开始从实验室走向市场。这一进步，很大程度上归功于深度学习的发展，深度学习在语音识别系统中的应用使得其性能显著提升。目前，语音识别正在朝着基于深度学习的端到端的方向发展。

### 11.1.1 流程框架

语音识别的过程就是将输入语音经过处理后，得到语音特征，通过模式匹配的方法，将输入语音与语音库中的语音特征进行匹配，将相似度高的结果作为输出。目前，对于语音识别任务，大部分都是基于隐马尔科夫模型（Hidden Markov Model，HMM）来实现的。音频输入在隐马尔科夫模型中是可观测变量，而对应的文本在模型中是隐变量。

为了更好表明问题，我们先做出一些定义。

1）音频输入O，它通常是通过对波形文件进行分帧处理得到的一系列声学特征向量：

2）句子W，它是由一系列单词或词向量组成：

有了以上定义，根据语音识别任务要求，给定音频输入O得到最有可能的句子，可以形式化地将语音识别任务表示为数学式：

根据贝叶斯公式得：

由于音频输入O是不变的，可将此公式简化为：

其中，为可观察序列的似然概率，由声学模型提供；为先验概率，由语言模型提供。

语音识别的系统主要有五部分组成：信号处理、特征提取、声学模型、语言模型、解码器，其结构框图如图11.1所示。但在基于深度学习的语音识别中，模糊了各阶段的处理过程，它企图使用一个深层神经网络完成特征提取、声学模型和语言模型的构建从而对语音特征解码，实现端到端的语音识别。



图11.1 语音识别系统的结构框图

其中，最关键的部分就是声学模型、语言模型部分，本节主要对声学模型进行了讲解，而语言模型将在下一章深度学习在自然语言处理中的应用中进行讲解，自然语言处理是它更重要的一个应用。

### 11.1.2 声学模型

#### 7.1.2.1 概述

声学模型的任务是给模型产生语音波形的概率。将声学和发音学的知识进行整合，以特征提取模块提取的特征为输入，生成声学模型得分。声学模型是语音识别系统的重要组成部分，它占据着语音识别大部分的计算开销，决定着语音识别系统的性能。

传统的语音识别系统普遍采用的是基于GMM-HMM的声学模型，其中GMM用于对语音声学特征的分布进行建模，HMM则用于对语音信号的时序性进行建模。2006年深度学习兴起以后，深度神经网络（Deep Neural Networks，DNN）被应用于语音声学模型。2009年，Hinton及其学生将前馈全连接深度神经网络应用于语音识别声学建模，在TIMIT数据库上基于DNN-HMM的声学模型相比于传统的GMM-HMM声学模型可以获得显著的性能提升。DNN相比于GMM的优势在于：1）DNN对语音声学特征的后验概率进行建模不需要对特征的分布进行去分布假设；2）GMM要求对输入的特征进行去相关处理，而DNN可以采用各种形式的输入特征；3）GMM只能采用单帧语音作为输入，而DNN则可以通过拼接相邻帧的方式利用上下文的有效信息。2011年，DengLi等提出基于CD-DNN-HMM的声学模型，在大词汇量连续语音识别任务上取得成功，相比于传统的GMM-HMM系统可以获得超过20%的相对性能提升。

基于DNN-HMM的语音声学模型开始取代GMM-HMM成为主流的声学模型。此后大量的研究人员投入到基于深度神经网络的语音声学建模研究中，语音识别取得了突破性的进展。

#### 7.1.2.2 主流算法

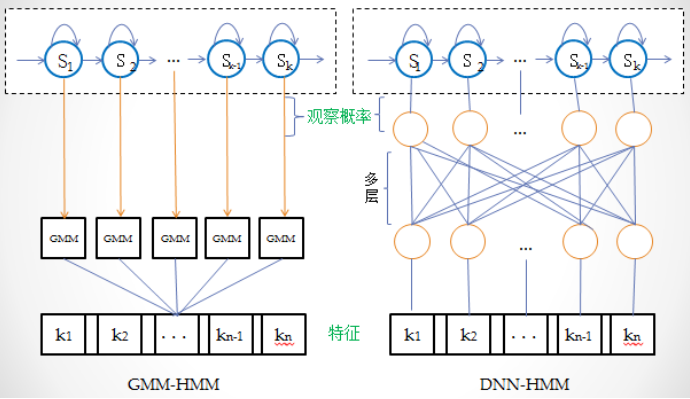


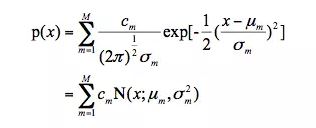
图11.2

**传统声学模型（GMM-HMM）**

HMM模型对时序信息进行建模，在给定HMM的一个状态后，GMM对属于该状态的语音特征向量的概率分布进行建模（也就是表示两者的关系）。

**1) 混合高斯模型**

如果一个连续随机变量服从混合高斯分布，则它的概率密度函数为：



混合高斯模型分布最明显的性质是它的多模态，这使得混合高斯模型可以描述很多显示出多模态性质的物理数据，比如语音数据，而单高斯分布则不合适。数据中的多模态性质可能来自多种潜在因素，每一个因素决定分布中特定的混合成分。如果因素被识别出来，那么混合分布就可以被分解成有多个因素独立分布的集合。

**2) 隐马尔科夫模型**

为了描述语音数据，在马尔可夫链的基础上进行了扩展，用一个观测的概率分布与马尔可夫链上的每个状态进行对应，这样引入双重随机性，使得马尔可夫链不能被直接观察，故称为隐马尔可夫模型。隐马尔可夫模型能够描述语音信号中不平稳但有规律可学习的空间变量。具体的来说，隐马尔可夫模型具有顺序排列的马尔可夫状态，使得模型能够分段的处理短时平稳的语音特征，并以此来逼近全局非平稳的语音特征序列。

优点：GMM训练速度快；

声学模型较小，容易移植到嵌入式平台；

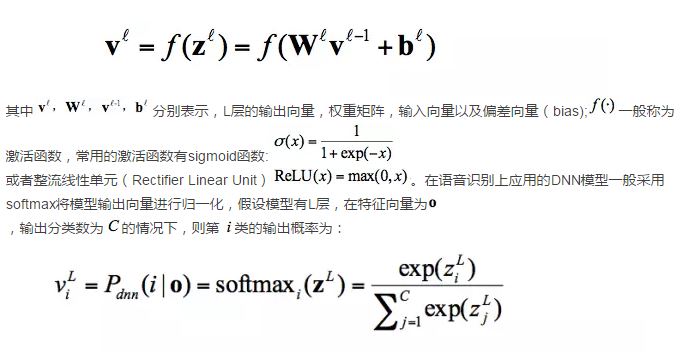
缺点：GMM没有利用帧的上下文信息；

GMM不能学习深层非线性特征变换。

**基于深度学习的声学模型（CD-DNN-HMM）**

虽然GMM-HMM在以往取得了很多成功，但是随着深度学习的发展，DNN模型展现出了明显超越GMM模型的性能，替代了GMM进行HMM状态建模。不同于GMM模型，DNN模型为了获得更好的性能提升，引入了上下文信息（也即前后特征帧信息），所以被称为CD-DNN-HMM（Context-Dependent DNN-HMM）模型。在很多测试集上CD-DNN-HMM模型都大幅度超越了GMM-HMM模型。

首先简单介绍一下DNN模型，DNN模型是有一个有很多隐层的多层感知机，下图就是具有5层的DNN，模型结构上包括输入层、隐层和输出层。对于第L层，有公式:



相比于GMM模型，DNN模型具有一些明显的优势：首先，DNN是一种判别模型，自身便带有区分性，可以更好区分标注类别；其次，DNN在大数据上有非常优异的表现，伴随着数据量的不断增加，GMM模型在2000小时左右便会出现性能的饱和，而DNN模型在数据量增加到1万小时以上时还能有性能的提升；另外，DNN模型有更强的对环境噪声的鲁棒性，通过加噪训练等方式，DNN模型在复杂环境下的识别性能甚至可以超过使用语音增强算法处理的GMM模型。

除此之外，DNN还有一些有趣的性质，比如，在一定程度上，随着DNN网络深度的增加，模型的性能会持续提升，说明DNN伴随模型深度的增加，可以提取更有表达性、更利于分类的特征；人们利用这一性质，提取DNN模型的Bottle-neck特征，然后在训练GMM-HMM模型，可以取得和DNN模型相当的语音识别效果。

DNN应用到语音识别领域后取得了非常明显的效果，DNN技术的成功，鼓舞着业内人员不断将新的深度学习工具应用到语音识别上，从CNN到RNN再到RNN与CTC的结合等等，伴随着这个过程，语音识别的性能也在持续提升，未来我们可以期望将可以和机器进行无障碍的对话。

### 11.1.3 应用与实例

## 11.2 声纹识别

声纹识别（Voiceprint），也称说话人识别（Speaker Recognition），是一种通过声音判别说话人身份的技术。声纹识别和语音识别在原理上一样，都是通过对采集到的语音信号进行分析和处理，提取相应的特征或建立相应的模型，然后据此做出判断。但二者的根本目的，提取的特征、建立的模型是不一样的。语音识别的目的是识别语音的内容；声纹识别的目的是识别说话人的身份。

根据识别场景不同，声纹识别分为两类：说话人识别（Speaker Identification，SI）、说话人确认（Speaker Verification，SV）；根据是否与说话内容有关，声纹识别又可分为：文本相关的声纹识别（Text-Dependent）、文本独立的声纹识别（Text-Independent）。

### 11.2.1 流程框架

声纹识别过程与语音识别过程类似，它主要有四部分组成，分别是信号处理、特征提取、声纹模型、判别器，其处理流程如图11.3所示。



图11.3 声纹识别处理流程

声纹识别系统也是一个典型的模式识别的框架，为了让计算机认识一个用户的声纹，需要用户首先提供一段语音，用于模型训练，最后会被映射为用户的声纹模型。在这之后，一个身份未知的音频输入会经过一系列的操作被映射为若干特征，通过此特征与声纹模型中特征进行相似度计算后得到一个置信度的得分，此置信度得分便是声纹识别产生的结果，它是评判这个身份未知的声音来源与已知声音来源的相似程度。因此，识别性能好坏的关键在于对语音中身份信息的建模能力与区分能力。

### 11.2.2 声纹模型

由于在声纹识别过程中，声纹模型起着主要作用，接下来我们主要对声纹模型做介绍。

#### 11.2.2.1 概述

声纹模型的发展大概分为七个阶段。

1. 模板匹配：此阶段只是基于信号的比对来建模；
2. 高斯混合模型（GMM）：此阶段需要用大量数据为每个说话人训练模型，注册模型需要录入很长时间的语音；
3. 高斯混合背景模型（GMM-UBM）及支持向量机（GMM-SVM）：虽然减少了建模注册时需要的数据量，但是对跨信道分析的能力较弱；
4. 联合因子分析（JFA）： 此阶段分别建模了说话人信道空间以及残差噪声，但每一步都会引入一定误差；
5. 基于GMM的I-Vector方法及PLDA：统一建模所有信道空间，进一步减少了模型注册时间，用PLDA分辨说话人特征，但噪声对GMM依然有很大影响；
6. 基于深度神经网络的I-Vector方法： DNN不仅可以从音频中学到大量高维特征，而且对噪声的鲁棒性也增强了；
7. 端到端的深度神经网络：由DNN直接端到端的实现声纹模型，不必分步依次进行特征提取、分类。

#### 11.2.2.2 主流算法

**传统声纹模型**

1. **GMM-UBM**

GMM是一种参数化的生成性模型，具备对实际数据极强的表征力。GMM中高斯分量的数量足够多的时候，GMM可以模拟任意的概率分布，但是与此同时也需要更多数据来训练模型。

DA Reynolds团队，提出了一种改进的方案。既然不能获取大量目标用户的的语音，那么可以通过其它手段收集大量非目标用户的语音。我们可以将这些非目标用户数据（背景数据）混合起来充分训练出一个GMM，这个GMM可以看作是对语音的表征。这个混合GMM类似于贝叶斯框架的先验模型，它对语音特征在空间分布的概率模型给出了一个良好的预先估计，我们不需再从头开始计算GMM的参数，只需要基于目标用户的数据在这个混合GMM上进行参数的微调即可实现目标用户参数的估计，这就是GMM-UBM模型。

GMM-UBM模型的一个重要的优势在于它是通过最大后验估计（MAP）的算法对模型参数进行估计，减少了过拟合程度；另外一个优势是不必再去调整目标用户GMM的所有参数只需要对各个高斯成分的均值参数进行估计。但是，由于参数较多，仍需要大量数据进行训练。此外，GMM-UBM模型缺乏对信道多变性的补偿能力，鲁棒性较低。

1. **GMM-SVM**

WM Campbell将支持向量机引入了GMM-UBM的模型中，对GMM中每个高斯分量均值构建一个高斯超向量（Gaussian Super Vector，GSV）作为SVM的样本，利用带核函数的SVM的非线性分类能力，在原始GMM-UBM的基础上大幅提升了识别性能。同时为了补偿由于信道易变形对声纹建模带来的影响，还使用了基于GSV的一些规整算法，如扰动属性投影（Nuisance Attribute Projection, NAP），类内方差规整（Within Class Covariance Normalization，WCCN）等。这在一定程度上加强模型对跨信道分析的能力。

**3）GMM-I-Vector**

Dehak提出了从GMM均值超矢量中提取一个更紧凑的矢量，称为I-Vector（Identity-Vector）。在I-Vector模型中，采用全局差异空间(Total Variability Space，T)，同时包含了说话人之间的差异和信道之间的差异。所以I-Vector的建模过程在GMM均值超矢量中不严格区分说话人的影响和信道的影响。

给定说话人的一段语音，这一新的说话人、信道相关的GMM均值超矢量定义为如下公式：

其中，m是说话人、信道无关的超矢量，即UBM超矢量；T是全局差异因子，它是一个低秩的矩阵；w就是I-Vector，它具有标准正态分布的随机向量，其分量数是总因子数。可知，M是服从均值为m、协方差矩阵为的正态分布。可以认为特定说话人，特定语音的M，由UBM的均值加全局空间矩阵与全局差异因子的矩阵乘积来确定。

在I-Vector的确认中，需要基于如下两个关键步骤：全局差异空间矩阵T的估计、I-Vector的估计。

对于全局差异空间矩阵T的估计，认为每一段语音都来自不同的说话人，采用如下过程来估计T矩阵：

1）计算训练数据库中每个说话人所对应的Baum-Welch统计量；

2）T的初始值，用EM算法迭代估计T矩阵，E步计算隐变量w的后验分布，M步进行最大似然值重估，重新更新T矩阵。

全局差异空间矩阵T估计完毕后，便可以获得I-Vector矢量，然后进行后续基于I-Vector的声纹模型。

**基于深度学习的声纹模型**

随着这几年深度学习的发展，越来越多的基于深度学习的声纹模型被提出，使得声纹模型的性能得到了提升。

一类是在传统模型上进行改进，使用深度学习来完成其中某一部分任务。Daniel Garcia-Romero et al.提出一种利用深度神经网络提高声纹识别的性能，他们使用DNN模型来代替GMM模型，使用经过域外数据训练后的DNN收集统计信息，可以显著提高识别性能，如图11.4所示。P. Kenny et al.提出了一种使用DNN来提取Baum-Welch统计量，不是使用标准EM算法训练通用背景模型，而是预定义组件并且对应于三元组状态，其后验用概率由DNN建模。



(a)基于GMM (b)基于DNN

图11.4 Daniel Garcia-Romero et al.提出的基于DNN的声纹模型

另一类是直接使用深度神经网络实现端到端的声纹模型。其基本网络结构如图11.5所示。



图11.5 基于深度学习端到端声纹模型的基本网络结构

DNN在声纹识别中的应用使得识别性能得到了显著提升。目前，基于CNN、RNN、LSTM等网络结构的声纹模型正在越来越多被提出，基于深度学习的声纹识别方法、声纹模型正在发展中。

### 11.2.3 应用与实例

## 参考文献

语音识别框架原理简述<https://blog.csdn.net/u012528143/article/details/51873467>

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups （2012年，Hinton，语音识别的突破）

Speech recognition with deep recurrent neural networks （2013年，RNN在语音识别的应用）

Towards End-To-End Speech Recognition with Recurrent Neural Networks （2014年）

Fast and accurate recurrent neural network acoustic models for speech recognition （2015年，Google）

Deep speech 2: End-to-end speech recognition in english and mandarin （2015年，Baidu）

Achieving Human Parity in Conversational Speech Recognition （2016年，Microsoft）

深度学习在语音识别中的算法、应用、数据集、行业分析<https://blog.csdn.net/qq_37572875/article/details/77899491>

语音识别、声纹识别的区别及测试<https://blog.csdn.net/alice_tl/article/details/82319781>

Deep Neural Networks for Speaker Recognition, Xingxing Tang <https://wenku.baidu.com/view/a8ab106d76232f60ddccda38376baf1ffc4fe3ae.html>

声纹识别综述<https://blog.csdn.net/Allyli0022/article/details/78195352>

SVM-GMM Based Speaker Identification

Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.

The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge

Deep neural networks for extracting baum-welch statistics for speaker recognition

IMPROVING SPEAKER RECOGNITION PERFORMANCE IN THE DOMAIN ADAPTATION CHALLENGE USING DEEP NEURAL NETWORKS

Deep Speaker: an End-to-End Neural Speaker Embedding System

End-to-End Text-Dependent Speaker Verification