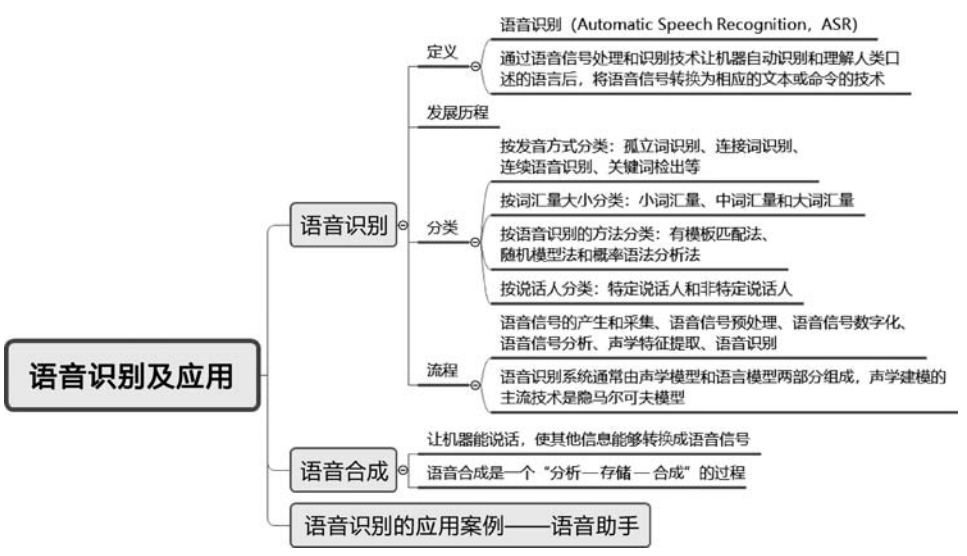
第 5 章

CHAPTER 5

语音识别及应用

本章思维导图



视频讲解

本章目标

- 了解语音识别的发展、概念。
- 理解语音识别的分类、基本原理。
- 理解语音合成的基本原理。

5.1 语音识别

语言是人类最重要的交流工具，而语音是语言的声学表现形式，是人类最自然的交互方式，具有准确高效、自然方便的特点。随着社会的发展，越来越多的机器参与到了人类的生产活动和社会活动中，因此改善人们和机器之间的关系，让人们对机器的操纵更加方便、灵活就显得越来越重要。随着人工智能的发展，人们发现，语音通信是人和机器之间最好的通

信方式，如图 5-1 所示。

图 5-1　语音通信


视频讲解

5.1.1　语音识别的定义

语音识别（Automatic Speech Recognition，ASR）是以语音为研究对象，通过语音信号处理和识别技术让机器自动识别和理解人类口述的语言后，将语音信号转变为相应的文本或命令的技术，使得人机用户界面更加自然和容易使用。语音识别技术是指，与机器进行语音交流，让机器明白人说什么，这是人们长期以来梦寐以求的事情。可以形象地把语音识别比作为“机器的听觉系统”、人工智能的“耳朵”。

语音识别是一门涉及面很广的交叉学科，与声学、语音学、语言学、信息理论、模式识别理论以及神经生物学等学科都有非常密切的关系。语音识别技术主要包括特征提取技术、模式匹配准则及模型训练技术 3 个方面。


视频讲解

语音识别技术已经在现实生活中得到了广泛应用，具有广阔的应用前景，如语音检索、命令控制、自动客户服务、机器自动翻译等。

语音识别技术的五大应用领域如下：

- 电信。话务员协助服务的自动化、国际国内远程电子商务、语音呼叫分配、语音拨号、分类订货。
- 医疗。由声音来生成和编辑专业的医疗报告、语音医疗记录等。
- 制造业。在质量控制中，语音识别系统可以为制造过程提供一种“不用手”“不用眼”的检控（部件检查），增加人与机器的语音交互界面，由语音对机器发出命令，机器用语音做出应答。
- 办公室或商务系统。填写数据表格、数据库管理和控制、键盘功能增强、语音会议记录等。
- 其他。由语音控制和操作的游戏和玩具，以及帮助残疾人的语音识别系统等。

当今信息社会的高速发展迫切需要性能优越的、能满足各种不同需求的自动语音识别技术。

5.1.2　语音识别发展历程

语音识别技术的目标是研究出一种具有听觉功能的机器，能够接收人类的语音，理解人的意图。由于语音识别本身所固有的难度，人们提出了各种限制条件下的研究任务，并由此产生了不同的研究领域。

1952 年，贝尔实验室的 Davis 等人研制出了特定说话人孤立数字识别系统。这个系统能够利用每个数字元音部分的频谱特征进行语音识别。1959 年，Fry 和 Denes 等人尝试构建音素识别器，用于识别 4 个元音和 9 个辅音，采用频谱分析和模式匹配来进行识别决策，其突出贡献在于，使用了英语音素序列中的统计信息来改进词中音素的精度。

20 世纪 60 年代初期，日本的研究者开发了相关的特殊硬件来进行语音识别，如东京无线电研究实验室研制的通过硬件来进行元音识别的系统。在此期间开展的很多研究工作对后来近二十年的语音识别研究产生了很大的影响。

20 世纪 70 年代以前，语音识别的研究特点是以孤立词的识别为主。20 世纪 70 年代，语音识别研究在很多方面取得了诸多成就。在孤立词识别方面，日本学者 Sakoe 给出了使用动态规划方法进行语音识别的途径——DTW 算法，这是语音识别中一种非常成功的匹配算法，当时在小词汇量的研究中获得了成功，从而掀起了语音识别的研究热潮。也是在这个时期，人工智能技术开始被引入到语音识别中。

20 世纪 80 年代，识别算法从模式匹配技术转向基于统计模型的技术，是语音识别研究的一个重要进展，更倾向于从整体统计的角度来建立最好的语音识别系统。隐马尔可夫模型(Hidden Markov Model，HMM)就是其中的一个典型，该模型被广泛地应用到语音识别研究中。到目前为止，HMM 模型仍然是语音识别研究中的主流方法。这些研究工作开创了语音识别的新时代。

从 20 世纪 80 年代后期和 90 年代初开始，人工神经网络(Artificial Neural Network，ANN)的研究异常活跃，并且被应用到语音识别的研究中。进入 20 世纪 90 年代后，相应的研究工作在模型设计的细化、参数提取和优化，以及系统的自适应技术等方面取得了一些关键性的进展，使语音识别技术进一步成熟，并且出现了一些很好的产品。

进入 21 世纪，基于深度学习理论的语音识别得到了全面突破，识别性能显著提高。随着深度学习技术的发展，卷积神经网络和循环神经网络等网络结构成功地应用到语音识别任务中，目前能够彻底摆脱 HMM 框架的端到端语音识别技术正日益成为语音识别研究的焦点，无论是学术机构，还是工业界都投入大量的人力和财力，致力于此方面的研究。

我国语音识别研究工作起步于 20 世纪 50 年代，近年来发展速度很快，研究水平也从实验室逐步走向实用。从 1986 年开始执行“国家 863”计划后，国家 863 智能计算机专家组为语音识别技术研究专门立项，每两年滚动一次。我国语音识别技术的研究水平已经基本上与国外同步，在汉语语音识别技术上还有自己的特点与优势，并达到国际先进水平。其中具有代表性的研究单位为清华大学电子工程系与中科院自动化研究所模式识别国家重点实验室。

清华大学电子工程系语音技术与专用芯片设计课题组，研发的非特定人汉语数码串连续语音识别系统的识别精度，达到 94.8%(不定长数字串)和 96.8%(定长数字串)。在有 5%拒识率的情况下，系统识别率可以达到 96.9%(不定长数字串)和 98.7%(定长数字串)，这是目前国际最好的识别结果之一，其性能已经接近实用水平。研发的 5000 词邮包校核非特定人连续语音识别系统的识别率达到 98.73%，前三项识别率达 99.96%；并且可以识别普通话与四川话两种语言，达到实用要求。

采用嵌入式芯片设计技术研发了语音识别专用芯片系统，该芯片以 8 位微控制器(MCU)核心，加上低通滤波器、模/数(A/D)、数/模(D/A)、功率放大器、RAM、ROM、脉宽

调幅(PWM)等模块，构成了一个完整的系统芯片，这是国内研发的第一块语音识别专用芯片。芯片中包括了语音识别、语音编码、语音合成功能，可以识别 30 条特定人语音命令，识别率超过 95%，其中的语音编码速率为 16kb/s。该芯片可以用于智能语音玩具；也可以与普通电话机相结合构成语音拨号电话机。这些系统的识别性能完全达到国际先进水平。研发的成果已经进入实用领域，一些应用型产品正在研发中，其商品化的过程也越来越快。


视频讲解

5.1.3 语音识别的分类

语音识别技术有多种不同的分类方法。

(1) 按发音方式进行分类，可以分为孤立词识别、连接词识别、连续语音识别、关键词检出等几种类型。在孤立词识别中，机器仅识别一个个孤立的音节、词或短语等，并给出具体识别结果；在连续语音识别中，机器识别连续自然的书面朗读形式的语音；而连接词识别中，发音方式介于孤立词和连续语音之间，它表面上看像连续语音发音，但能明显地感觉到音到音之间有停顿。这时通常可以采用孤立词识别的技术进行串接来实现；对关键词检出，通常用于说话人以类似自由交谈的方式发音，称为自发发音方式时；在这种发音方式下，存在着各种各样影响发音不流畅的因素，如犹豫、停顿、更正等，并且说话人发音中存在着大量的不是识别词中的词，判断理解说话人的意思，只从其中一些关键的部分就可做出决定，因此只需进行其中的关键词的识别。

(2) 按词汇量大小进行分类，每一个语音识别系统都有一个词汇表，语音识别系统只能识别出词汇表中所包含的词条。通常按词汇量大小可以分为小词汇量(一般包括 10～100 个词条)、中词汇量(一般包括 100～500 个词条)和大词汇量(包括 500 个以上的词条)3 类。通常情况下，随着语音识别系统中词汇量的增大，语音识别的识别率会降低，因此，在这种分类下语音识别的研究难度会随着词汇量的增多而增加。

(3) 按语音识别的方法进行分类，有模板匹配法、随机模型法和概率语法分析法。这些方法都属于统计模式识别方法。其识别过程大致如下：首先提取语音信号的特征构建参考模型，然后用一个可以衡量未知模板和参考模板之间似然度的测度函数，选用一种最佳准则和专家知识作出识别决策，给出识别结果。其中模板匹配法是将测试语音与参考模型的参数一一进行比较与匹配，判决的依据是失真测度最小准则。随机模型法是一种使用隐马尔可夫模型来对似然函数进行估计与判决，从而得到相应的识别结果的方法。由于隐马尔可夫模型具有状态函数，所以这个方法可以利用语音频谱的内在变化(如说话速度、不同说话人特性等)和它们的相关性。概率语法分析法适用于大范围的连续语音识别，它可以利用连续语音中的语法约束知识来对似然函数进行估计和判决。其中，语法可以用参数形式来表示，也可以用非参数形式来表示。

(4) 按说话人进行分类，可以分为特定说话人和非特定说话人两种。前者只能识别固定某个人的声音。其他人要想使用这样的系统，必须事先输入大量的语音数据，对系统进行训练；而对后者，机器能识别任意人的发音。由于语音信号的可变性很大，这种系统要能从大量的不同人(通常为 30～40 人)的发音样本中学习到非特定人的发音速度、语音强度、发音方式等基本特征，并归纳出其相似性作为识别的标准。使用者无论是否参加过训练都可以共用一套参考模板进行语音识别。从难度上看，特定说话人的语音识别比较简单，能得到较高的识别率，并且目前已经有商品化的产品；而非特定人识别系统，通用性好、应用面广，

但难度也比较大，不容易获得较高的识别率。

在语音识别中，最简单的是特定人、小词汇量、孤立词的语音识别，最复杂、最难解决的是非特定人、大词汇量、连续语音识别。无论是哪一种语音识别，当今采用的主流算法仍然是隐马尔可夫模型方法。

5.1.4　语音识别的流程


视频讲解

语音识别的流程分为语音信号的产生和采集、语音信号预处理、语音信号数字化、语音信号分析、声学特征提取、语音识别等，如图 5-2 所示。

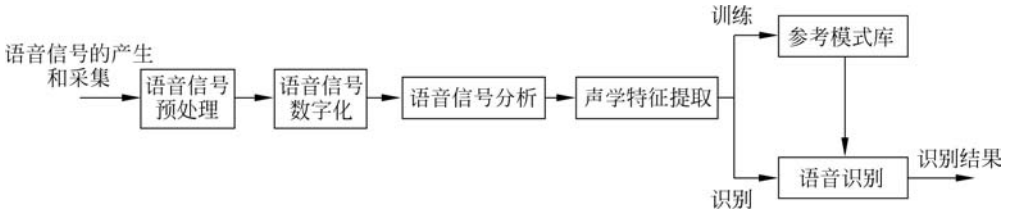


图 5-2　语音识别流程图

下面分别介绍语音识别流程中各部分的内容。

1. 语音信号的产生和采集

物理课上曾学习过声波的产生和传播原理——声波是由物体振动产生。说话人的发声器官做出发音动作，在空气中振动形成声波，通过空气传到听者的耳朵，最后到达人耳被人感知，声波由耳郭收集之后经一系列结构的传导到达耳蜗，耳蜗内有丰富的听觉感受器，可将声音传导到听神经，最后引起听者的听觉反应，语音的传递就是这样一个过程。人类耳朵的结构如图 5-3 所示。

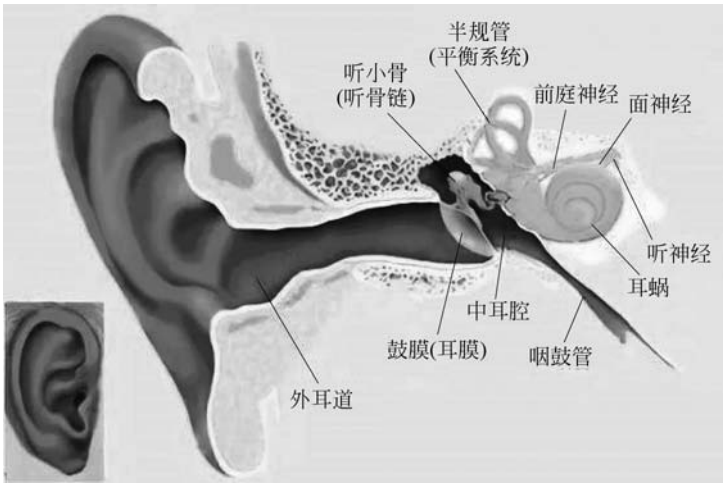


图 5-3　耳朵结构图

频率是声音的重要特征，代表了发生物体在一秒内振动的次数，单位是赫兹，人耳的精妙结构也决定了对不同频率的声音有着不同的敏感。如图 5-4 所示，横坐标代表频率，纵坐标代表引起人耳听觉的声音强度，单位是分贝，这个值越小代表人对频率的声音越敏感。声

音作为一种波，频率在 20Hz～20kHz 的声音是可以被人耳识别的。

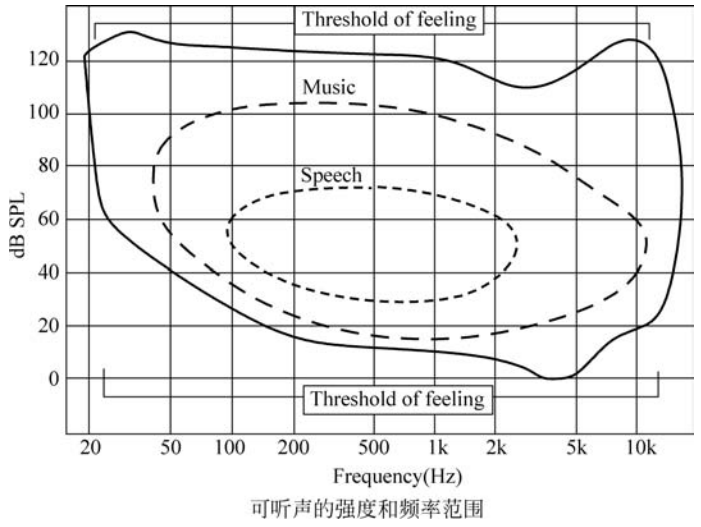


图 5-4　人的听觉频率范围

首先，说话人在头脑中产生想要用语言表达的信息，然后将这些信息转换成语言编码，即将这些信息用其所包含的音素（音素是指发出各不相同音的最小单位）序列、韵律、响度、基音周期的升降等表示出来。一旦这些信息编码完成后，说话人就会用一些神经肌肉命令在适当的时候控制声带振动，并塑造声道的形状以便可以发出编码中指定的声音序列。神经肌肉命令必须同时控制调音运动中涉及的各个部位，包括唇、舌头等，以及控制气流是否进入鼻腔的软腭。

语音是以声波的方式在空气中传播。语音信号一旦产生，并传递到计算机时，计算机通过话筒对语音信息进行采集。话筒将声波转换为电压信号，然后通过 A/D 装置（如声卡）进行采样，从而将连续的电压信号转换为计算机能够处理的数字信号。模数转换器即 A/D 转换器，是指一个将模拟信号转变为数字信号的电子元件。一般是将一个输入电压信号转换为一个输出的数字信号。

2. 语音信号预处理

语音信号数字化之前，必须先进行预处理，包括防混叠滤波及防工频干扰滤波。在得到的声波信号输入中需要实际处理的信号并不一定占满整个时域（一个信号的时域波形可以表达信号随着时间的变化），会有静音和噪声的存在，因此，必须先对得到的输入信号进行一定的预处理，进行防混叠滤波和防工频干扰滤波，其中防混叠滤波是指滤除高于 1/2 采样频率的信号成分或噪声，使信号带宽限制在某个范围内，否则如果采样率不满足采样定理，则会产生频谱混叠，此时信号中的高频成分将产生失真，而工频干扰是指 50Hz 的电源干扰。

3. 语音信号数字化

如何让计算机感知声音，这时候就需要将声波转换为便于计算机存储和处理的音频文件（比如 mp3 文件），这个过程如图 5-5 所示，从声波到最终的 mp3 文件主要经历了采样、离散化、量化和编码等步骤。

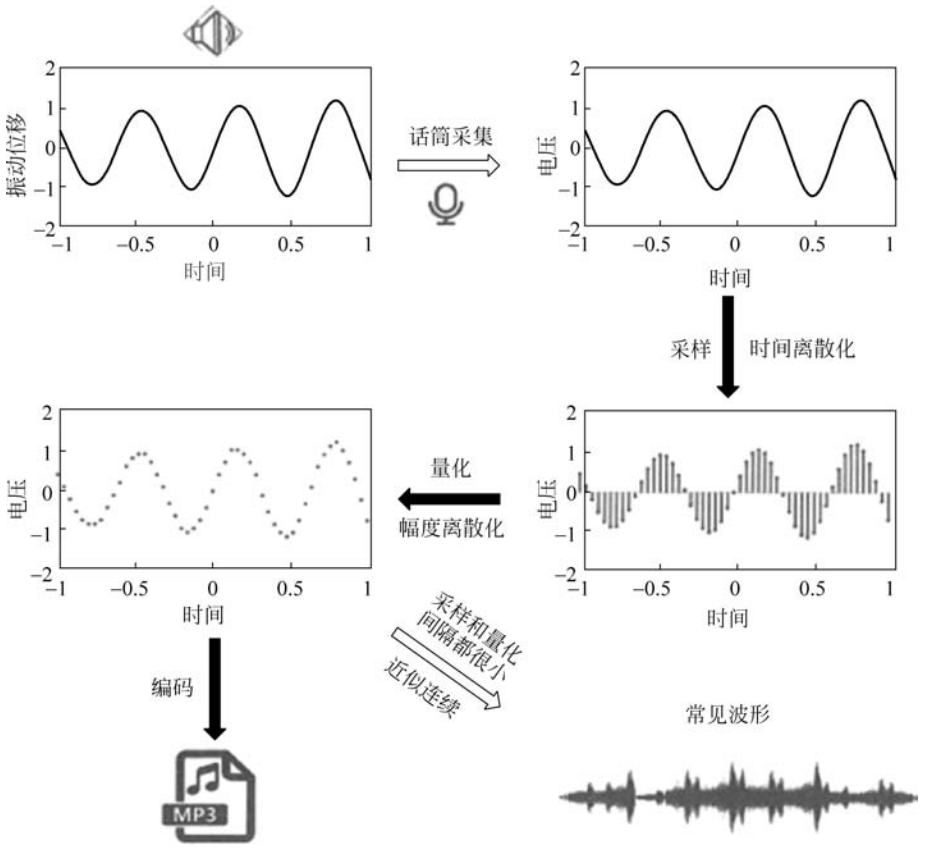


图 5-5　声音的数字化

语音信号是时间和幅度都连续变化的一维模拟信号，要想在计算机中对模拟信号进行处理，就要先进行采样和量化，将其变成时间和幅度都离散的数字信号。现今得到广泛应用的音频文件格式（如 mp3 等）都经过了压缩而无法直接识别。语音识别所使用的音频文件格式必须是未经压缩处理的 wav 格式文件。计算机里面的音频文件描述的实际上是一系列按时间先后顺序排列的数据点，所以也称为时间序列。把时间序列可视化出来就是常见的波形，其横坐标表示时间，纵坐标没有直接的物理意义，反映了传感器在传导声音时的振动位移。

在进行语音信号数字处理时，最先接触、最直观的是它的时域波形。通常是将语音用话筒转换成电信号，再用模数转换器将其转换成离散的数字采样信号后，存入计算机中。由于数字信号本身不具有实际意义，仅仅表示一个相对大小。因此，任何一个模数转换器都需要一个参考模拟量作为转换的标准，比较常见的参考标准为最大的可转换信号大小，而输出的数字量则表示输入信号相对于参考信号的大小。

经常采样和量化过程后，一般还要对语音信号进行一些预加重。由于语音信号的平均功率谱受声门激励和口鼻辐射的影响，高频端大约在 800Hz 以上按着 −6dB/倍频程跌落，为此要在预处理中进行预加重。其目的就是提升高频部分，使信号的频谱变得平坦，便于进行频谱分析或声道参数分析，如图 5-6 所示。

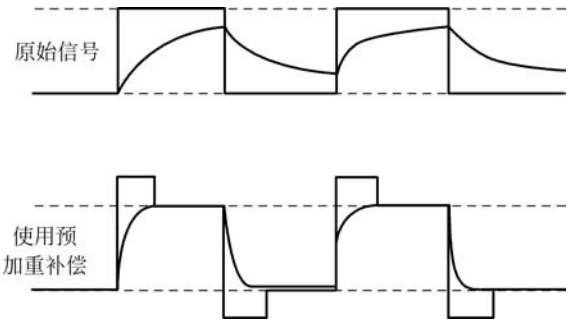


图 5-6　预加重补偿

数字化之后的语音信号还需要进行一些处理工作，最常用的前端处理有端点检测和语音增强。端点检测是指在语音信号中将语音和非语音信号时段区分开来，准确地确定出语音信号的起始点。经过端点检测后，后续处理就可以只对语音信号进行，这对提高模型的精确度和识别正确率有重要作用。语音增强的主要任务就是消除环境噪声对语音的影响。目前通用的方法是采用维纳滤波，该方法在噪声较大的情况下效果好于其他滤波器。从连续的（或离散的）输入数据中滤除噪声和干扰以提取有用信息的过程称为滤波，这是信号处理中经常采用的主要方法之一，具有十分重要的应用价值，相应的装置称为滤波器。

只有消去静音的部分并且滤除噪声的干扰，才能使处理后的信号更能反映语音的本质特征，才能对实际需要处理的有效语音进行识别，所以在开始语音识别之前，通常需要把首尾端的静音切除，降低对后续步骤造成的干扰。

4. 语音信号分析

语音是一种特殊的声音，所以它具有声学特征的物理性质。语音的声学特征是指音色、音高、音长和音强，简称语音的四要素。音色也称为音质，是一种声音区别于其他声音的基本特征。从物理学角度来分析，音调的变化其实对应频率的变化，也就是基频随着声调的变化而变化。

语音识别的前提是对语音信号的分析。只有将语音信号分析表示成其本质特性的参数，才有可能利用这些参数进行高效的语音通信，才能建立用于语音合成的语音库，也才可能建立用于识别的模板或知识库。而且，语音合成的音质好坏、语音识别率的高低，都取决于对语音信号分析的准确性和精度。所以，应该先对语音信号进行特征分析，得到提高语音识别率的有用数据，并据此来设计语音识别系统的硬件和软件。

语音分析的工作必须先于其他的语音信号处理工作。根据所分析的参数不同，语音信号分析可以分为时域、频域、倒谱域等方法。进行语音信号分析时，最先接触到的、最直观的是它的时域波形。语音信号本身就是时域信号，因而时域分析是最早使用且应用范围最广的一种方法。时域分析具有简单直观、清晰易懂、运算量小、物理意义明确等优点，但更为有效的分析多是围绕频域进行的，因为语音中最重要的感知特性反映在其功率谱中，而相位变化只起着很小的作用。

根据语音学的观点，可将语音信号分析分为模型分析法和非模型分析法两种。模型分析法是指依据语音信号产生的数学模型，来分析和提取表征这些模型的特征参数；共振峰模型分析及线性预测分析即属于这种方法。凡不进行模型化分析的其他方法都属于非模型

分析法，包括时域分析法、频域分析法及同态分析法等。

贯穿于语音信号分析全过程的是“短时分析技术”。根据对语音信号的研究，其特性是随时间而变化的，所以它是一个非稳态过程。从另一方面看，虽然语音信号具有时变特性，但不同的语音是由人的口腔肌肉运动构成声道的某种形状而产生的声响，而这种肌肉运动频率相对于语音频率来说是缓慢的，因而在短时间范围内，其特性基本保持不变，即相对稳定，所以可以将其看作是一个准稳态过程。基于这样的考虑，对语音信号的分析和处理必须建立在“短时”的基础上，即进行“短时分析”。将语音信号分为一段一段来分析，其中每一段称为一“帧”。由于语音信号通常在 10～30ms 是保持相对平稳的，因而帧长一般取 10～30ms，如图 5-7 所示。

图 5-7　分帧

取出来的一帧信号，在做傅里叶变换之前，要先进行“加窗”的操作，即与一个“窗函数”相乘，如图 5-8 所示。数字化仪器采集到的有限序列的边界会呈现不连续性。加窗可减少这些不连续部分的幅值。加窗包括将时间记录乘以有限长度的窗，窗的幅值逐渐变小，在边沿处为 0。加窗的结果是尽可能呈现出一个连续的波形，减少剧烈的变化。

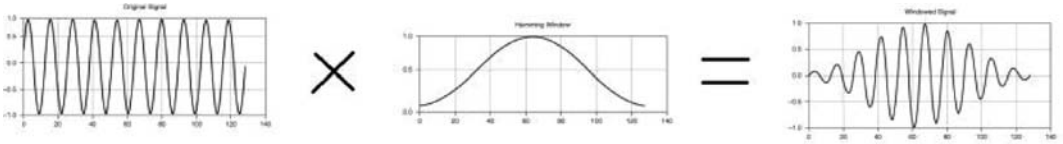


图 5-8　加窗

加窗的目的是让一帧信号的幅度在两端渐变到 0。渐变对傅里叶变换有好处，可以提高变换结果（即频谱）的分辨率。加窗的代价是一帧信号两端的部分被削弱了，没有像中央的部分那样得到重视。弥补的办法是，帧不要背靠背地截取，而是相互重叠一部分。相邻两帧的起始位置的时间差叫作帧移，常见的取法是取为帧长的一半，或者固定取为 10ms。否则，由于帧与帧连接处的信号会因为加窗而被弱化，这部分的信息就丢失了。

对一帧信号做傅里叶变换，得到的结果叫频谱（一般只保留幅度谱，丢弃相位谱），如图 5-9 所示的蓝线，该图中的横轴是频率，纵轴是幅度。频谱上就能看出这帧语音在 480Hz 和 580Hz 附近的能量比较强。语音的频谱，常常呈现出“精细结构（音高）”和“包络（音素）”两种模式。“精细结构”就是蓝线上的一个个小峰，它们在横轴上的间距就是基频，它体现了语音的音高——峰越稀疏，基频越高，音高也越高。“包络”则是连接这些小峰峰顶的平滑曲线（红线），它代表了口型，即发的是哪个音。包络上的峰叫共振峰（共振峰是指声音频谱上能量相对集中的一些区域），图中能看出 4 个，分别在 500Hz、1700Hz、2450Hz、3800Hz 附近。有经验的人，根据共振峰的位置，就能看出发的是什么音。

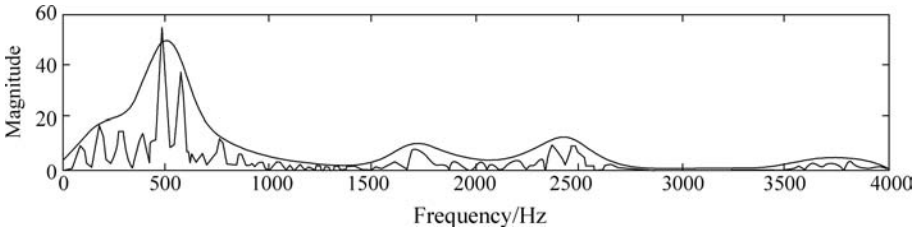


图 5-9　傅里叶变换后的频谱图（见彩插）

5. 声学特征提取

声学特征的提取与选择是语音识别的一个重要环节。模拟的语音信号进行采样得到波形数据之后，首先要送到特征提取模块，提取出合适的声学特征参数，供后续声学模型训练使用。好的声学特征应当考虑以下 3 个方面的因素。首先，应当具有比较优秀的区分特性，以使声学模型不同的建模单元可以方便准确地建模；其次，特征提取也可以认为是语音信息的压缩编码过程，既需要将信道、说话人的因素消除，保留与内容相关的信息，又需要在不损失过多有用信息的情况下使用尽量低的参数维度，便于高效准确地进行模型的训练；最后，需要考虑鲁棒性，也就是对环境噪声的抗干扰能力。

经过数字化的语音信号实际上是一个时变信号，这是由于人在发音时声道一直处于变化状态，因此实际上的语音信号产生系统可以近似看作线性时变系统。典型的语音信号特性是随着时间变化而变化的。例如，浊音和清音之间激励的改变，会使信号峰值幅度有很大的变化，在浊音范围内基频有相当大的变化。在一个语音信号的波形图中，这些变化十分明显，所以要求能用简单的时域处理技术来对这样的信号特征进行有效的描述。

在语音识别和说话人识别中，常用的语音特征是基于 Mel 频率的倒谱系数（Mel Frequency Cepstrum Coefficient，MFCC），是在 Mel 标度频率域提取出来的倒谱参数，Mel 标度描述了人耳频率的非线性特性，一定程度上模拟了人耳对语音的处理特点，MFCC 所处的位置如图 5-10 所示。由于 MFCC 参数是将人耳的听觉感知特性和语音的产生机制相结合，因此大多数语音识别系统都使用了这种特征。MFCC 特征是基于人耳对声音的敏感特性而提出的常用的一种方法，叫作梅尔频率倒谱系数，通过 L 维（L 可以取值为 12～16）的向量来描述一帧的波形，L 维向量是根据耳朵的生理特征提取的，这一过程称为声学特征提取。声音就被转换成了 L 行 N 列的矩阵（观察序列）。

前面做完傅里叶变换之后，接下来把频谱与图 5-11 中每个三角形相乘并积分，求出频谱在每一个三角形下的能量。将能量谱通过一组 Mel 尺度的三角形滤波器组，定义一个有 M 个滤波器的滤波器组（滤波器的个数和临界带的个数相近），采用的滤波器为三角滤波器，第 m 个滤波器的中心频率为 $f(m)$。M 通常取 22～26。各 $f(m)$之间的间隔随着 m 值的减小而缩小，随着 m 值的增大而增宽，如图 5-11 所示。

三角带通滤波器有两个主要目的：

（1）对频谱进行平滑化，并消除谐波的作用，突显原先语音的共振峰（因此一段语音的音调或音高，是不会呈现在 MFCC 参数内，换句话说，以 MFCC 为特征的语音辨识系统，并不会受到输入语音的音调不同而有所影响）。

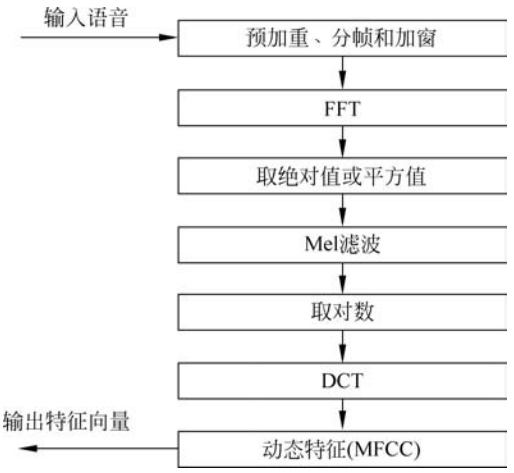


图 5-10　MFCC 架构图

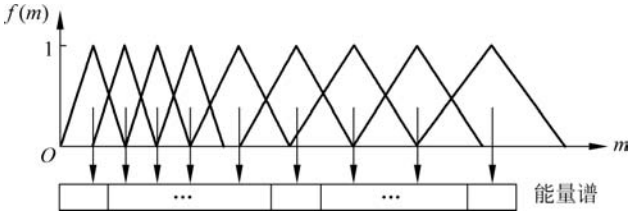


图 5-11　Mel 频率滤波器组

（2）此外，还可以只保留需要的信息，降低运算量。如此一来，就把一帧语音信号用一个 L 维向量简洁地表示了出来；一整段语音信号，就被表示为这种向量的一个序列。这时候，语音就可以通过一系列的倒谱向量来描述了，每个向量就是每帧的 MFCC 特征向量，如图 5-12 所示。语音识别中下面要做的事情，就是对这些向量及它们的序列进行建模了。

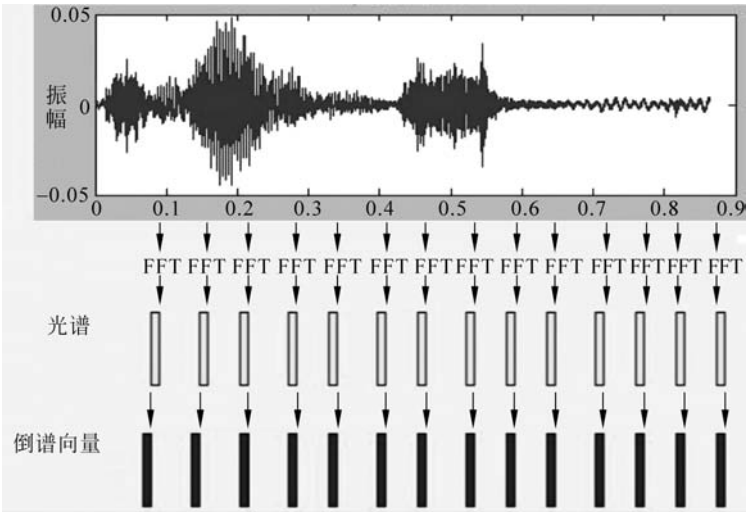


图 5-12　MFCC 特征提取

MFCC 是一组特征向量，反映了频谱的轮廓，可用于音色分类。对人耳听觉机理的研究发现，人耳对不同频率的声波有不同的听觉敏感度。200～5000Hz 的语音信号对语音的清晰度影响最大。两个响度不等的声音作用于人耳时，则响度较高的频率成分的存在会影响到对响度较低的频率成分的感受，使其变得不易察觉，这种现象称为掩蔽效应。一般来说，低音容易掩蔽高音，而高音掩蔽低音较困难。在低频处的声音掩蔽的临界带宽较高频要小。所以，人们从低频到高频这一段频带内按临界带宽的大小由密到疏安排一组带通滤波器，对输入信号进行滤波。将每个带通滤波器输出的信号能量作为信号的基本特征，对此特征经过进一步处理后就可以作为语音的输入特征。由于这种特征不依赖于信号的性质，对输入信号不做任何假设和限制，从而利用了听觉模型的研究成果。因此，这种参数具有更好的鲁棒性，更符合人耳的听觉特性，而且当信噪比降低时仍然具有较好的识别性能。

6. 语音识别

将待识别的语音经特征提取后，逐一与参考模板库中的各个模板按某种原则进行比较，找出最相像的参考模板所对应的发音，即为识别结果。

语音识别系统的模型通常由声学模型和语言模型两部分组成，分别对应于语音到音节概率的计算和音节到字概率的计算，如图 5-13 所示。音素一般就是熟知的声母和韵母，而状态则是比音素更加细节的语音单位，把帧识别成状态，把状态组合成音素，把音素组合成单词。每帧音素对应哪个状态，看某帧对应哪个状态的概率最大，这帧就属于哪个状态。这些概率可以从声学模型里读取，里面存了许多参数，通过这些参数，就可以知道帧和状态对应的概率。获取这一大堆参数的方法叫作“训练”。一个音素通常会包含 3 个状态(起始音、持续音、结束音)，把一系列语音帧转换为若干音素的过程利用了语音的声学特性，因此这部分叫作声学模型。隐马尔可夫模型是目前进行声学建模的主流技术。从音素到文字的过程需要用到语言表达的特点，这样才能从同音字中挑选出正确的文字，组成意义明确的语句，这部分被称为语言模型。

语音识别系统的前提是需要建立参考模板库。在训练阶段，对特征参数形式表示的语音信号进行相应的技术处理，获得表示识别基本单元共性特点的标准数据，以此来构成参考模板，参考模板库是由所有能识别的基本单元的参考模板综合在一起形成的。

如图 5-14 所示，语音识别系统中的模型训练分为声学模型训练和语言模型训练两部分。

1) 声学模型训练

声学模型训练也称为建模的过程。声学模型是语音识别系统的底层模型，是语音识别系统中最关键的部分。声学模型表示一种语言的发音，可以通过训练来识别某个特定用户的语音模式和发音环境的特征。根据训练语音库的特征参数训练出声学模型参数，在识别时可以将待识别的语音的特征参数同声学模型进行匹配与比较，得到最佳识别结果。

基本声学单元的选择是声学模型建模中一个基本而重要的问题。在汉语连续语音识别中，可以选择的基本声学单元包括词、音节、半音节、声韵母、音素等。

一般来说，声学单元越小，其数量也就越少，训练模型的工作量也就越小，但是另一方面，单元越小，对上下文的敏感性越大，越容易受到前后相邻的影响而产生变异，因此其类型设计和训练样本的采集就更加困难。通常要根据不同的语音识别系统有针对性地选择适合的基本声学单元。其中，声韵母是适合汉语特点建模的基本声学单元。

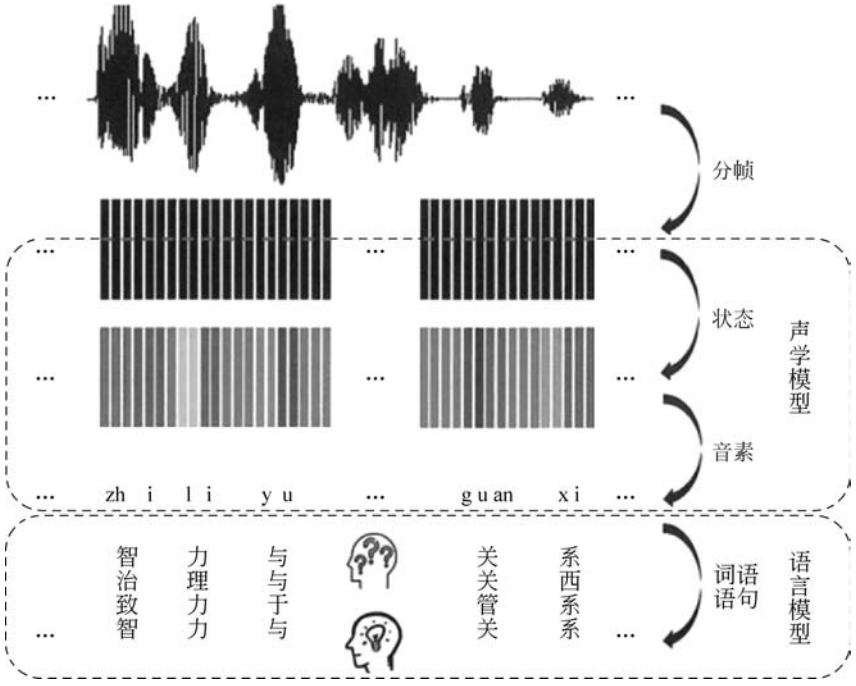


图 5-13　语音识别

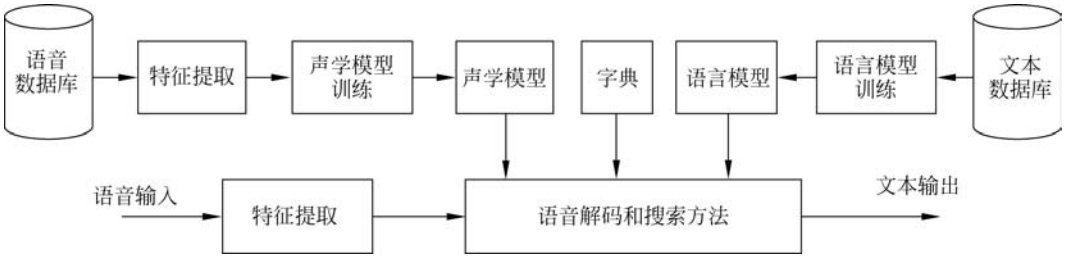


图 5-14　模型训练示意图

为了得到满意、有效的模型，必须有很多训练数据。目前主流的训练技术是隐马尔可夫模型。

2）语言模型训练

语音识别中的语言模型主要解决两个问题：一是如何使用数学模型来描述语音中词的语音结构；二是如何结合给定的语言结构和模式识别器形成识别算法。语言模型是用来计算一个句子出现概率的概率模型，主要用于决定哪个词序列的可能性更大，或者在出现了几个词的情况下预测下一个即将出现的词语的内容。换句话说，语言模型是用来约束单词搜索的，定义了哪些词能跟在上一个已经识别的词的后面（匹配是一个顺序的处理过程），这样就可以为匹配过程排除一些不可能的单词。语言模型一般指在匹配搜索时用于字词和路径约束的语言规则，它包括由识别语音命令构成的语法网络或由统计方法构成的语言模型，语言处理则可以进行语法和语义分析。

语言建模能够有效地结合汉语语法和语义的知识，描述词之间的内在关系，从而提高识

别率，减少搜索范围。语言模型分为 3 个层次：字典知识、语法知识、句法知识。对训练文本数据库进行语法和语义分析，经过基于统计模型训练得到语言模型。

前面介绍的语音识别系统均由多个模块组成，一般包括声学模型、语言模型、发音词典等。其中声学模型和语言模型需要分别独立训练得到，它们各自有不同的目标函数。

另外，目前语音识别一个热门的研究方向是端到端的语音识别技术。近年来，研究者正在探索端到端的语音识别技术，它试图用一个神经网络来承担原来所有模块的功能。这样，系统中将不再有多个独立的模块，而仅通过神经网络来实现从输入端（语音波形或特征序列）到输出端（单词、音素或音符的序列）的直接映射。端到端的识别技术能有效减少人工预处理和后续处理，避免了分阶段学习问题，能给模型提供更多的基于数据驱动的自动调节空间，从而有助于提高模型的整体契合度。

5.2 语音合成

语音合成的主要目的是让机器能说话，以便使一些其他存储方式的信息能够转化成语音信号，让人能够简单地通过听觉获得大量的信息。语音合成技术除了在人机交互中的应用外，在自动控制、测控通信系统、办公自动化、信息管理系统、智能机器人等领域也有广阔的应用前景。

语言合成应用场景如下：

- 阅读类 App——通过阅读类 App 阅读小说或新闻时，使用语音合成技术为用户提供多种发音人的朗读功能，获得更好的阅读体验。
- 订单播报——可应用于打车软件、餐饮叫号、排队软件等场景，通过语音合成进行订单播报，让用户便捷地获得通知信息。
- 智能硬件——可集成到儿童故事机、早教机、智能机器人、平板设备等智能硬件设备，使用户与设备的交互更自然、更亲切。
- 呼叫中心——为满足各行业声音需求，企业可以根据自身品牌需求定制个性化音色服务。
- 其他——合成特定人的声音，验证码内容语言合成，各场景的语言提示（如导航软件、大厅、售货机），便携式穿戴设备（播报每日的健康指数）。

目前各种语音报警器、语音报时器、公共汽车上的自动报站、股票信息的查询、电话查询业务等均已实现商品化。另外，语音合成技术还可以作为听觉、视觉和语音表达有障碍的伤残人士的通信辅助工具。

语音合成是一个“分析-存储-合成”的过程。一般是选择合适的基元，将基元用一定的参数编码方式或波形方式进行存储，形成一个语音库。合成时，根据待合成的语音信息，从语音库中取出相应的基元进行拼接，并将其还原成语音信号。在语音合成中，为了便于存储，必须先将语音信号进行分析或变换，因而在合成前还必须进行相应的反变换。其中，基元是语音合成系统所处理的最小的语音学基本单元，待合成词语的语音库就是所有合成基元的集合。根据基元的选择方式及其存储形式的不同，可以将合成方式笼统地分成波形合成方法和参数合成方法。

（1）波形合成方法是一种相对简单的语音合成技术。它把人的发音波形直接存储或者

进行简单波形编码后存储，组成一个合成语音库；合成时，根据待合成的信息，在语音库中取出相应单元的波形数据，拼接或编辑到一起，经过解码还原成语音。这种系统中的语音合成器主要完成语音的存储和回放任务。如果选择如词组或者句子这样较大的合成单元，则能够合成高质量的语句，并且合成的自然度好，但所需要的存储空间也相当大。虽然在波形合成法中，可以使用波形编码技术压缩一些存储量，但由于存储容量的限制，词汇量不可能做到很大。通常，波形合成法可合成的语音词汇量约在500字以下，一般以语句、短句、词或者音节为合成基元。

(2) 参数合成方法也称为分析合成方法，是一种比较复杂的方法。为了减少存储空间，必须先对语音信号进行各种分析，用有限个参数表示语音信号以压缩存储容量。参数的具体表示，可以根据语音生成模型得到诸如线性预测系统、线谱对参数或共振峰参数等。这些参数比较规范、存储量少。参数合成方法的系统结构较为复杂，并且用参数合成时，由于在提取参数或编码过程中，难免存在逼近误差，用有限个参数很难适应语音的细微变化，所以合成的语音质量以及清晰度也就比波形合成法要差一些。

就目前的技术水平，仅采用上述的“分析-存储-合成”的思想是不可能合成任一语种的无限词汇量的语音。因而国际上很多研究者都在努力开发另一类无限词汇量的语音合成的方法，就是所谓“按语言学规则的从文本到语言”的语音合成法，简称“规则合成方法”。人们期望通过这项研究合成出高自然度的语音来，尽管到目前为止还未曾获得这样的效果。

5.3　语音识别的应用案例——语音助手

训练已经构建好的声学模型，并保存模型文件，最后整合声学模型和语言模型，实现中文语音识别。其中，声学模型GRU-CTC，只需导入路径即可：

```
from model_speech.gru_ctc import Am, am_hparams
```

实现此案例需要按照如下步骤进行。

【步骤1】　声学模型训练

编辑train.py文件完成声学模型的训练，语言模型使用已经训练好的模型文件。

【案例5-1】　train.py

```
import keras
import os
import tensorflow as tf
from utils import get_data, data_hparams
from keras.callbacks import ModelCheckpoint


# 0.准备训练所需数据------------------------------


data_args = data_hparams()
data_args.data_type = 'train'
```

```
data_args.data_path = '../dataset/'
data_args.thchs30 = True
data_args.aishell = True
data_args.prime = True
data_args.stcmd = True
data_args.batch_size = 4
data_args.data_length = 10
# data_args.data_length = None
data_args.shuffle = True
train_data = get_data(data_args)

# 0.准备验证所需数据------------------------------
data_args = data_hparams()
data_args.data_type = 'dev'
data_args.data_path = '../dataset/'
data_args.thchs30 = True
data_args.aishell = True
data_args.prime = False
data_args.stcmd = False
data_args.batch_size = 4
# data_args.data_length = None
data_args.data_length = 10
data_args.shuffle = True
dev_data = get_data(data_args)

# 1.声学模型训练-----------------------------------
from model_speech.gru_ctc import Am, am_hparams
am_args = am_hparams()
am_args.vocab_size = len(train_data.am_vocab)
am_args.gpu_nums = 1
am_args.lr = 0.0008
am_args.is_training = True
am = Am(am_args)

if os.path.exists('logs_am/model.h5'):
    print('load acoustic model...')
    am.ctc_model.load_weights('logs_am/model.h5')

epochs = 10
batch_num = len(train_data.wav_lst) // train_data.batch_size

# checkpoint
ckpt = "model_{epoch:02d}-{val_acc:.2f}.hdf5"
checkpoint = ModelCheckpoint(os.path.join('./checkpoint', ckpt), monitor='val_loss', save_
weights_only=False, verbose=1, save_best_only=True)

batch = train_data.get_am_batch()
dev_batch = dev_data.get_am_batch()
```

```
am.ctc_model.fit_generator(batch, steps_per_epoch = batch_num, epochs = 10, callbacks =
[checkpoint], workers = 1, use_multiprocessing = False, validation_data = dev_batch,
validation_steps = 200)
am.ctc_model.save_weights('logs_am/model.h5')
```

【步骤 2】 模型测试

在 test.py 文件中整合声学模型和语言模型。

【案例 5-2】 test.py

本案例中 test.py 文件完整代码如下：

```
import os
import difflib
import tensorflow as tf
import numpy as np
from utils import decode_ctc, GetEditDistance


# 0.准备解码所需字典,参数需和训练一致,也可以将字典保存到本地,直接进行读取
from utils import get_data, data_hparams
data_args = data_hparams()
train_data = get_data(data_args)


# 1.声学模型----------------------------------
from model_speech.gru_ctc import Am, am_hparams

am_args = am_hparams()
am_args.vocab_size = len(train_data.am_vocab)
am = Am(am_args)
print('loading acoustic model...')
am.ctc_model.load_weights('logs_am/model.h5')

# 2.语言模型-------------------------------------------
from model_language.transformer import Lm, lm_hparams

lm_args = lm_hparams()
lm_args.input_vocab_size = len(train_data.pny_vocab)
lm_args.label_vocab_size = len(train_data.han_vocab)
lm_args.Dropout_rate = 0.
print('loading language model...')
lm = Lm(lm_args)
sess = tf.Session(graph = lm.graph)
with lm.graph.as_default():
    saver = tf.train.Saver()
with sess.as_default():
    latest = tf.train.latest_checkpoint('logs_lm')
    saver.restore(sess, latest)
```

```
# 3.准备测试所需数据，不必和训练数据一致，通过设置 data_args.data_type 测试，
#此处应设为'test'，我用了'train'因为演示模型较小，如果使用'test'看不出效果，
#且会出现未出现的词.
data_args.data_type = 'train'
data_args.shuffle = False
data_args.batch_size = 1
test_data = get_data(data_args)

# 4.进行测试 ------------------------------------------
am_batch = test_data.get_am_batch()
word_num = 0
word_error_num = 0
for i in range(10):
    print('\n the ', i, 'th example.')
    #载入训练好的模型，并进行识别
    inputs, _ = next(am_batch)
    x = inputs['the_inputs']
    y = test_data.pny_lst[i]
    result = am.model.predict(x, steps = 1)
    #将数字结果转化为文本结果
    _, text = decode_ctc(result, train_data.am_vocab)
    text = ''.join(text)
    print('文本结果:', text)
    print('原文结果:', ''.join(y))
    with sess.as_default():
        text = text.strip('\n').split(' ')
        x = np.array([train_data.pny_vocab.index(pny) for pny in text])
        x = x.reshape(1, -1)
        preds = sess.run(lm.preds, {lm.x: x})
        label = test_data.han_lst[i]
        got = ''.join(train_data.han_vocab[idx] for idx in preds[0])
        print('原文汉字:', label)
        print('识别结果:', got)
        word_error_num += min(len(label), GetEditDistance(label, got))
        word_num += len(label)
print('词错误率:', word_error_num / word_num)
sess.close()
```

本章总结

- 语音识别（Automatic Speech Recognition，ASR）是以语音为研究对象，通过语音信号处理和识别技术让机器自动识别和理解人类口述的语言后，将语音信号转变为相应的文本或命令的技术。
- 语音识别的流程分为语音信号的产生和采集、语音信号预处理、语音信号数字化、语音信号分析、声学特征提取、语音识别等。
- 语音合成的主要目的是让机器能说话，以便使一些以其他方式存储的信息能够转化

成语音信号，让人能够简单地通过听觉就可以获得大量的信息。

- 语音合成是一个“分析-存储-合成”的过程。

本章习题

1. 下面关于语音识别的说法正确的是(　　)。
   A. 语音识别汉语的时候，只需要考虑拼音的发音就可以了
   B. 声音的特征提取和图像的特征提取一样，直接提取就可以了
   C. 隐马尔可夫模型是用来建立声学模型的一种技术
   D. 以上说法都是错误的
2. 下列属于语音识别技术应用的是(　　)。
   A. 语音播报　　B. 语音识别
   C. 音乐分类　　D. 音乐检索
3. 简述语音识别的基本流程。
4. 语音信号数字化之前为什么要进行预处理？
5. 语音识别和语音合成技术有什么不同？
6. 请分享一下你在生活中所见到的语音识别的应用。