1. 语音处理

**导读**

语音信号是人类进行交流的主要途径之一。由于语言和语音与人的智力活动密切相关，与社会文化和进步紧密相连，所以它具有最大的信息容量和最高的只能水平。利用现代手段研究语音处理技术，能够让我们更加有效地产生、传输、存储、获取和应用语音信息。语音处理不仅在通信、工业、国防和金融等领域有着广阔的应用场景，而且正在这件改变人机交互的方式。

**学习目标**

本章将从以下几个部分介绍语音处理：主要包括语音的基本概念、语音识别、语音合成、语音增强、语音转换和情感语音等。

**8.1 语音的基本概念**

人类通过发音器官发出的话语便是语音，它具有一定意义，可用来进行社会社交。语音是一种声音，具有声学特征的物理特性。根据发音方式的不同，可以将语音分为元音和辅音，辅音又可以根据声带有无振动分为清辅音和浊辅音。人可以感觉到频率在20Hz-30kHz、强度为-5dB -130dB的声音信号，因此，在音频处理中，在这范围以外的音频可以忽略。

构成语音的四要素为：音高、音强、音长、音色。音高指声波频率，即每秒钟振动次数的多少；音强指声波振幅的大小；音长指声波振动持续时间的长短，也称为“时长”；音色指声音的特色和本质，也称为“音质”。

语音经过采样后，在计算机中以波形文件的方式进行存储，这些波形文件反映了语音在时域上的变化，通过波形可以判断语音音强（或振幅）、音长等参数的变化，但我们很难从波形中分辨出不同的语音内容或不同的说话人。为了更好地反映不同语音的内容或音色差别，需要对语音进行频域上的转换，即提取语音频域的参数。常见的语音频域参数包括傅里叶谱、梅尔频率倒谱系数等。通过对语音进行离散傅里叶变换可以得到傅里叶谱，在此基础上将语音信号在频域上划分成不同子带，进而得到梅尔频率倒谱系数。梅尔频率倒谱系数是一种能够近似反映人耳听觉特点的频域参数，在语音识别和说话人识别上被广泛使用。

**8.2 语音识别**

语音识别是指将语音自动转换为文字的过程。利用语音识别技术，能让机器把语音信号转变为相应的文本或命令，让机器听懂人类的语音。



语音识别技术研究始于20世纪50年代初期，迄今位置已有六十多年的历史。19520年，贝尔实验室研制了世界上第一个能识别十个英文数字的识别系统。20世纪60年代最具代表的研究成果是基于动态时间规整的模板匹配方法，这种方法有效地解决了特定说话人孤立词语音识别中语速不均和不等长匹配的问题。20世纪80年代以后，基于隐马尔科夫模型的统计建模方法逐渐取代了基于模板匹配的方法，基于高斯混合模型-隐马尔科夫模型的混合声学建模技术推动了语音识别技术的蓬勃发展，最具代表的是英国剑桥大学的隐马尔可夫工具包（HTK）。2010年后，语音识别技术也随着深度神经网络的兴起和分布式计算技术的进步而不断提升。

语音识别系统主要分为四个部分：特征提取（信号处理）、声学模型、语言模型和解码搜索。语音识别系统的框架如图8.1所示。

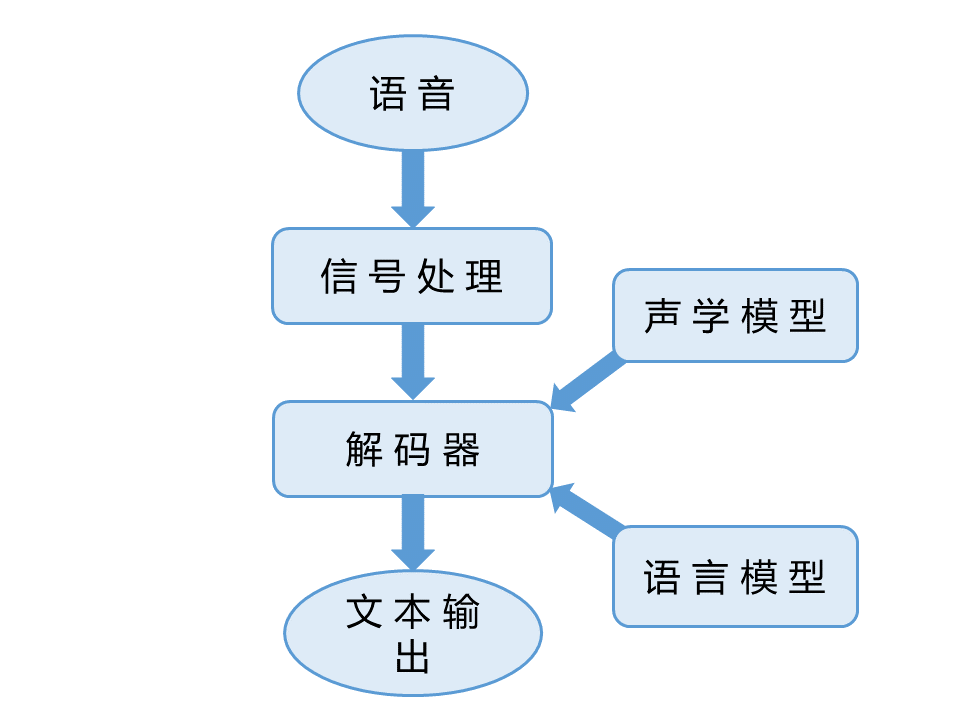


图8.1 语音识别系统框架

**8.2.1 语音识别的特征提取**

语音识别的难点之一在于语音信号的复杂性和多变性。语音特征提取即在原始语音信号中提取出语音识别最相关的信息，过滤其他无关信息。比较常用的声学特征有三种，即梅尔频率倒谱系数、梅尔标度滤波器组特征和感知线性预测倒谱系数。

梅尔频率倒谱系数特征是根据人耳听觉特性计算梅尔频谱域倒谱系数获得的参数。

梅尔标度滤波器组特征则是保留特征维度间的相关性。

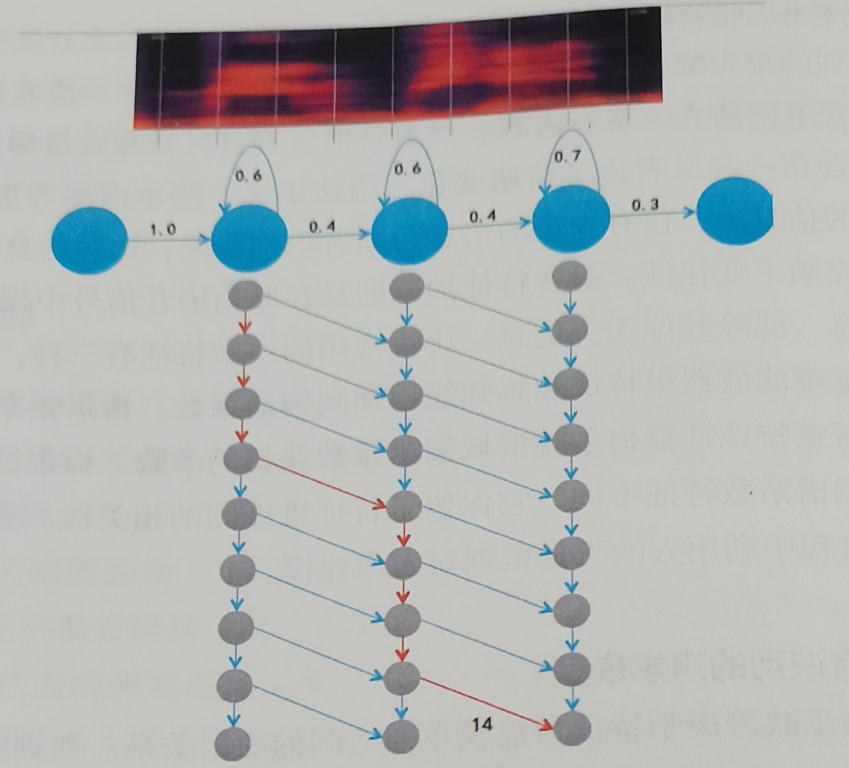
感知线性预测倒谱系数在提取过程中利用人的听觉机理对人声建模。

**8.2.2 语音识别的声学模型**

声学模型承载着声学特征与建模单元之间的映射关系。训练声学模型之前要选取建模单元。大多数声学模型一般采用音素作为建模单元，因为音素是上下文相关的。我们一般采用三音素进行声学建模。比较经典的声学模型是混合声学模型，大致分为两种：基于高斯混合模型-隐马尔科夫模型的模型(GMM-HMM)和基于深度神经网络-隐马尔科夫模型的模型(DNN-HMM)。

1. **基于高斯混合模型-隐马尔科夫模型的模型**

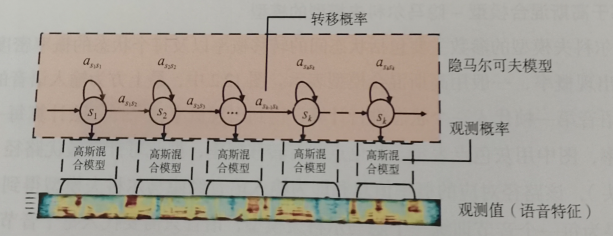
隐马尔科夫模型是一种概率图模型，可以用来表示序列之间的相关关系，经常被用来对时序数据进行建模。模型主要参数包括状态间的转移概率以及每个状态的概率魔都函数，也叫出现概率，一般用高斯混合模型表示。图8.2中，最上方为输入语音的音谱图，将语音第一帧带入一个状态计算能够得到出现概率；同样方法计算每一帧的出现出现概率，图中用灰色点表示。灰色点间有转移概率，据此可计算最优路径，如图8.2红色箭头所示，该路径对应的概率值总和即为输入语音经隐马尔科夫模型得到的概率值。如果为每个音节训练一个隐马尔科夫模型，只需按以上步骤，哪个得到概率最高即判定为相应音节，这也是传统语音识别的方法。



8.2隐马尔科夫模型示意图[1]

出现概率采用高斯混合模型，具有训练速度快、模型小、易于移植到嵌入平台等优点，缺点是没有利用帧的上下文信息，缺乏深层非线性特征变化的内容。高斯混合模型仅代表的一种概率密度，不能完整模拟出或记住相同音的不同人间的音色差异变化或发音习惯变化。

就基于GMM-HMM的声学模型而言，对于小词汇量的自动语音识别任务，通常使用上下文无关的音素状态作为建模单元；对于中等和大词汇的自动语音识别任务，则使用上下文相关的音素状态进行建模。该声学模型框架图如图8.3所示，高斯混合模型用来估计观察特征（语音特征）的观测概率，而隐马尔科夫模型责备用于描述语音信号的动态变化（即状态间的转移概率）。Sk代表音素状态；as1s2代表转移概率，即状态S1转为状态S2的概率。



8.3基于高斯混合模型-隐马尔科夫模型的声学模型[1]

1. **基于深度神经网络-隐马尔科夫模型的模型**

与GMM-HMM不同的是，DNN-HMM用深度神经网络模型代替了高斯混合模型。该模型的建模单元为聚类后的三音素状态，其框架如图8.4所示。图中，神经网络用来估计观察特征（语音特征）的观测概率。Sk代表音素状态；as1s2代表转移概率，即状态S1转为状态S2的概率；v代表输入特征；h(M)代表第M个隐层；WM代表神经网络第M个隐层的权重。

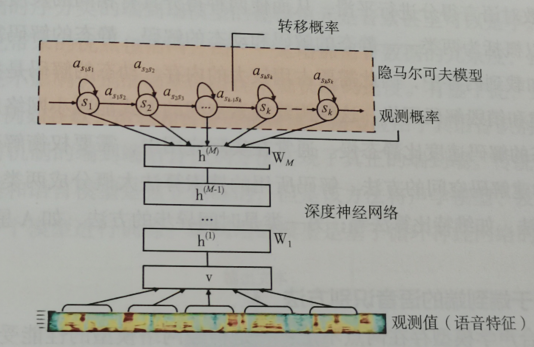


图8.4 基于深度神经网络 - 隐马尔科夫模型的声学模型[1]

这种模型有两方面的优势：一是深度神经网络能利用语音特征的上下文信息；而是深度神经网络能学习非线性的更高层次特征表达。因此DNN-HMM已成为目前主流的声学建模技术。

**8.2.3 语音识别的语言模型**

语言模型是根据语言客观事实而进行的语言抽象数学建模，可以表示某一字序列发生的概率。语言模型亦是一个概率分布模型P，用于计算任何句子S的概率。

例1：令句子S=“今天天气很好”，该句子很常见，通过语言模型可计算出其发生概率P（今天天气很好）=0.80000 。

例2：令句子S=“好很天气今天”，该句子为病句，通过语言模型可计算出其发生概率P（好很天气今天）=0.00001 。

语言模型的作用是在解码过程中限制搜索路径。语音识别中常用的语言模型是 N 元文法（N-Gram），即统计前后 N 个字出现的概率，以及循环神经网络语言模型。虽然循环神经网络语言模型性能由于N元文法，但训练比较耗时，且解码速度较慢，工业界依旧使用基于N元文法的语言模型。

语言模型的评价指标是语言模型在测试集上的困惑度，该值反映句子不确定性程度。因此我们的目标就是寻找困惑度较小的语言模型，使其尽量逼近真实语言分布

**8.2.4 语音识别的解码搜索**

解码搜索的主要任务是由声学模型、发音词典和语音模型构成的搜索空间中寻找最佳路径。解码时需要用到声学得分和语言得分，声学得分是由声学模型计算得到，语言得分是由语言模型计算得到。

在解码过程中，各种解码器的具体实现可以是不同的。按搜索空间的构成方式来分，有动态编译和静态编译两种方式。关于静态编译，是把所有知识源统一编译在一个状态网络中，在解码过程中，根据节点间的转移权重获得概率信息。就动态编译而言，只是预先将发音词典编译成状态网络构成搜索空间，其他知识源在解码过程中根据活跃路径上携带的历史信息动态集成。动态编译能够减小网络所站内存，但是解码速度比静态编译慢。

按搜索算法的时间模式来分，有异步与同步两种方式。时间异步的搜索算法通过栈解码器（Stack Decoder）来实现。时间同步的方法有Viterbi解码。基于树拷贝的帧同步解码器是目前比较流行的方法。

**8.2.5 基于端到端的语音识别方法**

上述混合声学模型仍存在着不足：（1）神经网络模型的性能受限于GMM-HMM模型的精度；（2）训练过程过于繁复。为了解决这些不足，研究人员提出端到端的语音识别方法，一类是基于联结时序分类的端到端声学建模方法；另一类是基于注意力机制的端到端语音识别方法。前者只是实现了声学建模的端到端，后者实现了真正意义上的端到端语音识别。

基于联结时序分类的端到端声学建模方法的声学模型结构如图8.5所示。这种方法只是在声学模型训练过程中，其核心思想是引入了一种新的训练准则联结时序分类。这种损失函数的优化目标是输入和输出在句子级别对齐，而不是帧级别对齐，因此不需要GMM-HMM生成强制对齐信息，而是直接对输入特征序列到输出单元序列的映射关系建模，极大简化了声学模型训练的过程。但是仍需对语言模型进行单独训练，从而构建解码的搜索空间。联结时序分类损失函数一般与长短记忆模型结合使用，基于联结时序分类的端到端模型的建模单元是音素甚至可以是字。这种建模单元粒度的变化带来的有点包括两方面：一是增加语音数据的冗余度，挺高音素的区分度；二是在不影响识别准确率的情况下加快解码速度。这种方法已被谷歌、微软和百度应用与其语音识别系统中。

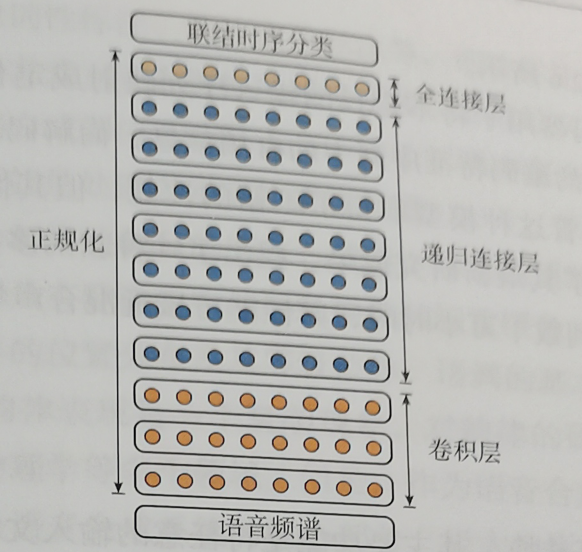


图8.5 基于连接时序分类的端到端声学模型结构图[1]

基于注意力机制的端到端语音识别方法实现了真正的端到端，该方法将声学模型、发音词典和语言模型联合为一个模型进行训练。模型基于循环神经网络的编码-解码结构，其结构如图8.6所示。解码器用于将不定长的输入序列映射成定长的特征序列，注意力机制用于提取编码器的编码特征序列中有用信息，而解码器则将该定长序列扩展成输出单元序列。尽管该模型取得不错的性能，但仍远不如混合声学模型。谷歌提出一种新的多头注意力机制的端到端模型，当训练数据达到数十万小时时，其性能可接近混合声学模型的性能。

图8.6 基于注意力机制的端到端语音识别系统结构图

**8.3 语音合成**

语音合成也称为文娱转换，其主要功能是将任意文本转换成语音的输出，这是人与计算机语音交互必不可少的模块。从地图导航，语音助手，小说、新闻朗读，智能音箱，语音实时翻译，到各种客服，机场广播，都少不了语音合成技术的身影。

图8.7所示是一个基本的语音合成框图。语音合成系统可以以任意文本作为输入，并相应地合成语音作为输出。语音合成系统主要分为文本分析模块、韵律处理模块和升学处理模块，其中文本分析模块可以视为系统的前端，二韵律处理模块和声学处理模块则视为系统的后端。

图8.7 语音合成系统框图

前端部分主要是对输入文本分析，从输入的文本提取后端建模需要的信息。例如：分词（判断句子中的单词边界），词性标注（名词，动词，形容词等），韵律结构预测（是否韵律短语边界），多音字消岐等等。后端的部分读入前端文本分析结果，并且对语音部分结合文本信息进行建模。在合成过程中，后端会利用输入的文本信息和训练好的声学模型，生成出语音信号，进行输出。

对于汉语拼音合成系统，文本分析的处理流程通常包括文本预处理、文本规范化、自动分词、词性标注、多音字消歧、节奏预测等，如图8.8所示。文本预处理包括删除无效符号、断句等。文本规范化的任务是将文本中的这些特殊字符识别出来并转化为一种规范化表达。自动分词是将待合成的整句以词尾单位划分为单元序列，一边后续考虑词性标注、韵律边界标注等。词性标注也很重要，因为词性可能影响字或词的发音。字音转换是将待合成的文字序列转换为对应的拼音序列。汉语存在多音字问题，所以字音转换的关键问题就是处理多音字的消歧问题。

图8.8 文本分析流程

韵律处理是文本分析模块的目的所在，节奏、时长的预测都基于文本分析的结果。韵律是实际语流中的抑扬顿挫和轻重缓急。作为语音合成系统中承上启下的模块，韵律模块实际是语音合成系统的核心部分，极大影响着最终合成语音的自然度。与韵律有关的语音参数包括基频、时长、停顿和能量，韵律模型利用文本分析结果预测这四个参数。

声学处理模块根据以上两个模块信息生成自然语言波形，现阶段的语音合成系统，根据所采用的方法和框架不同可以分为两种，一种是基于时域波形的拼接合成法，另一种是基于语音参数的合成法。

**8.3.1 基于拼接的语音合成方法**

基于拼接的语音合成方法的基本原理是将原始录音剪切成一个个基元存储下来，根据文本分析的结果，从预先录制并标注好的语音库中挑选合适基元进行适度调整，最终拼接得到合成语音波形。以上的基元是指语音拼接时的基本单元，可以时音节或者音素等。

拼接语音合成的优势在于，音质好，不受语音单元参数化的音质损失。但是在语料库小的情况下，由于有时挑选不到合适的语音单元，导致合成语音会有点问题 或者韵律、发音不够稳定，而且需要的存储空间大。大语料库则具有较高的的上下文覆盖率，挑选出来的基元几乎不需要任何调整就可以用于拼接合成，但稳定性仍然不够，可能出现拼接点不连续，以及难以改变发音特征。

**8.3.2基于参数的语音合成方法**

参数语音合成系统的基本思想是基于统计建模和机器学习的方法，特点是在语音分析阶段，需要根据语音生成的特点，将语音波形(speech waves) 通过声码器转换成频谱，基频，时长等语音或者韵律参数。在建模阶段对语音参数进行建模。并且在语音合成阶段，通过声码器从预测出来的语音参数还原出时域语音信号。

其中最成功的是基于隐马尔科夫模型的可训练语音合成方法，相应的合成系统被称为隐马尔科夫模型的参数合成系统。该主要包括训练阶段和合成阶段。图8.9展示的是该合成方法的系统框图。

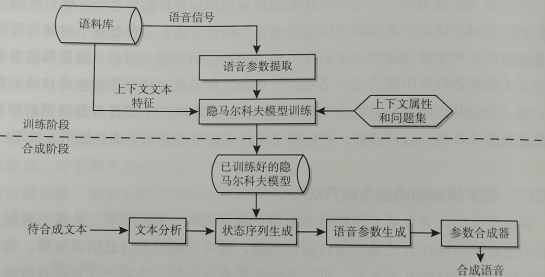


图8.9 基于隐马尔科夫模型语音合成系统框图[1]

在训练之前，需要对一些建模参数进行配置，包括建模单元的尺度、模型拓扑结构、状态数目等，还需要进行数据准备。一般而言，训练数据包括语音数据和标注数据两部分：标注数据主要包括音段切分和韵律标注。训练模型前还有一个重要的工作就是对上下文属性集和用于决策树聚类的问题进行设计，这部分工作是与语种相关的。其余步骤基本上与语言种类无关。

参数语音合成系统的优势在于模型大小较小，模型参数调整方便（说话人转换，升降掉），而且合成语音比较稳定。缺点在于合成语音音质由于经过参数化，所以和原始录音相比有一定的损失。

随着深度学习的发展，深度神经网络也被引入统计参数语音合成中，代替隐马尔科夫模型，可直接通过一个深层神经网络来预测声学参数，克服了隐马尔科夫模型训练中决策树聚类环节中模型精度降低的缺陷，进一步增强合成语音的质量。

**8.3.3 基于端到端的语音合成方法**

传统的语音合成流程很复杂。比如统计参数合成系统中通常会包含文本分析前端、时长模型、声学模型和基于复杂信号处理的声码器等模块，这些部分设计需要不同领域的知识，需要耗费大量精力，而且还需要分别训练，这意味着来自每个模块的错误可能会叠加。

2016年，谷歌Deep Mind提出一种基于深度学习的WaveNet 波形统计语音合成结构，主要单元是卷积神经网络，这种方法的特点是不会对语音信号进行参数化，而是用神经网络直接在时域预测合成语音波形的每一个采样点。优势是音质比参数合成系统好，略差于拼接合成。但是较拼接合成系统更稳定。缺点在于，由于需要预测每一个采样点，需要很大的运算量，合成时间慢。该模型仍然需要对来自现有语言合成文本分析前端的语言特征进行调节，并不是真正意义上端到端语音合成方法。

真正端到端工作是Tacotron语音合成系统。Tacotron是谷歌于2017年提出的端到端语音合成系统，该模型可接收字符的输入，输出相应的原始频谱图，然后将其提供给 Griffin-Lim 重建算法直接生成语音。该框架主要是基于注意力机制的编码-解码模型。其中，编码器是一个以字符或者音素为输入的神经网络模型；而解码器则是一个带有注意力机制的循环神经网络，会输出对应文本序列或者音素序列的频谱图，进而生成语音。这种语音合成方法的自然都和表现力已经能够媲美人类说话水平，而且不需要多阶段建模，现已成为当下热点和未来发展趋势。

**8.4 语音增强**

语音增强本质就是语音降噪，换句话说，日常生活中，麦克风采集的语音通常是带有不同噪声的“污染”语音，语音增强的主要目的就是从这些被“污染”的带噪语音中恢复出我们想要的干净语音。通过语音增强有效抑制各种干扰信号、增强目标语音信号，使人机之间更自然地交互。语音增强涉及的应用领域十分广泛，包括语音通话、电话会议、场景录音、军事窃听、助听器设备和语音识别设备等，并成为许多语音编码和识别系统的预处理模块。语音增强主要包括回声消除、混响抑制、语音降噪等关键技术。

**8.4.1 回声消除**

回声是指自身发出的声音经过多次反射（天花板，墙）并多次传入拾音设备。回声消除需要解决两个关键问题：①远端信号和近端信号的同步问题；②双讲模式下消除回波信号干扰的有效方法。回声消除最典型的应用是在智能终端播放音乐时，通过扬声器播放的音乐会回传给麦克风，此时便需要有效的回声消除算法抑制回声干扰。回声消除算法虽然提供了扬声器信号作为参考源，但是由于扬声器放音时的非线性失真、声音在传输过程中衰减、噪声干扰和回声干扰的同时存在，使得回声消除问题具有一定的挑战。

**8.4.2 混响抑制**

与回声不同的是，回声是声音发出结束后听到的声音，混响在声音还没结束听到的声音。适度的混响作用而使音乐更加动听。但在许多场合, 混响往往会带来干扰, 导致声学接收系统性能变差。房间大小、声源和麦克风的位置、室内障碍物、混响时间音素均影响混响语音的生成。

按照使用传声器数量分类, 去混响系统主要分为单传声器系统与传声器阵列系统。

单传声器系统去混响技术只利用声场中接收位置一点的声信号时间和变换域的特性。而多传声器阵列系统能利用声场的空间特性, 其主要优点是由阵列带来的接收方向性除了能直接提高信号与混响声能比之外, 同时还对本底噪声有显著的抑制作用。但阵列系统的硬件复杂度高, 数据处理量成倍增加, 对计算速度有较高要求。但随着计算机技术的发展, 采用阵列的去混响技术受到更多重视。

**8.4.3语音降噪**

噪声抑制可以分为基于单通道的语音降噪和基于多通道的语音降噪，前者通过单个麦克风去除各种噪声的干扰，后者通过麦克风阵列算法增强目标方向的声音。

基于单通道的语音降噪具有广泛的应用，在智能家居、智能客服、智能终端中均是非常重要的模块。单通道语音降噪有三种主流方法：①基于信号处理技术的语音降噪方法，该方法在处理平稳噪声具有不错性能，但面对非平稳噪声和突变噪声性能会下降；②基于矩阵分解的语音降噪方法，该方法计算复杂度相对较高；③基于数据驱动的语音降噪方法，当训练集和测试集不匹配时性能明显下降。

多通道语音降噪的目的时融合多个通道的信息，抑制非目标方向的干扰源，怎强目标方向的声音。需要解决的核心问题时估计空间滤波器，它的输入时麦克风阵列采集的多通道语音信号，输出时处理后的单路语音信号。多通道语音降噪算法还受限于麦克风阵列的结构，典型的阵列有线阵和环阵，阵列的选型与具体的应用场景相关。随着麦克风个数的增加，噪声抑制能力会更强，但算法复杂度和硬件功耗也会相应增加。

随着深度学习技术的快速发展，基于深度学习的语音降噪得到越来越广泛的应用，深层结构模型具有更强的泛化能力，在处理非平稳噪声具有更为明显的又是，这类方法更容易与语音识别的声学模型对接，提高语音识别的鲁棒性。

**8.5 语音转换**

语音转换指将一个人（源说话人）的声音个性化特征（如频谱、韵律等）通过“修改变换”，使之听起来像另外一个人（目标说话人）的声音，同时保持说话内容信息不变。广义上把改变语音中说话人个性特征的语音处理技术统称为语音转换。

语音转换首先需要提取说话人身份相关的声学特征参数，然后用改变后的声学特征参数合成出接近目标说话人的语音，如图8.10所示，语音转换系统包括训练阶段和转换阶段。训练阶段需要提取源说话人和目标说话人的个性特征参数，然后进行映射特征计算，进行转换模型训练；转换阶段根据训练获得的匹配函数对源说话人的个性特征参数进行转换，合成接近目标说话人的语音。

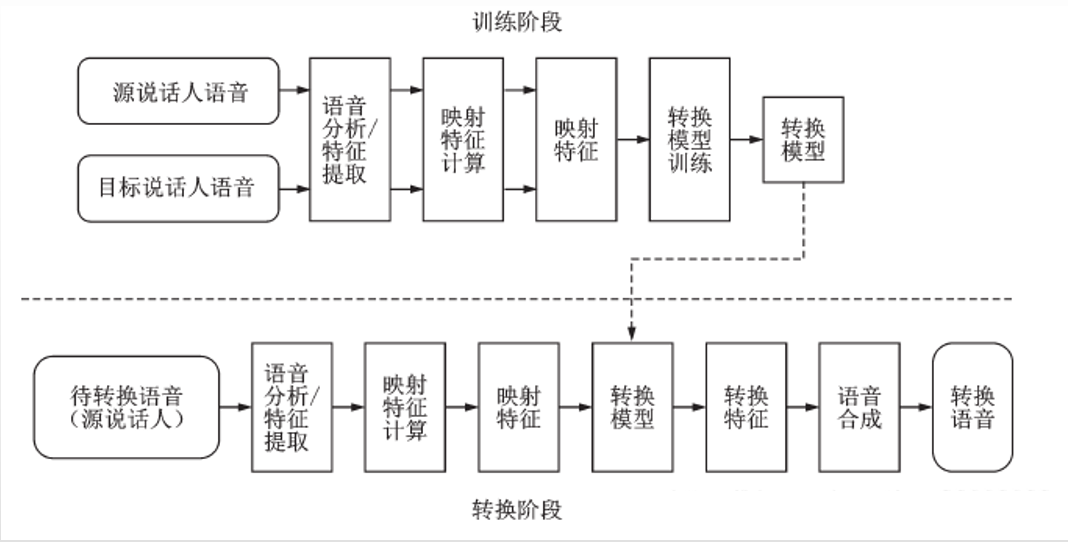


图8.10 语音转换基本系统框图

**8.5.1码本映射法**

码本映射法时最早应用于语音拼接的方法，这是一种比较有效的频谱转换算法。在这个算法中，源码本和目标码本的单元一一对应，通过从原始语音片段中抽取关键的语音帧作为码本，建立起源说话人和目标说话人参数空间的关系。

码本映射法的优点在于，由于码本从原始语音片段中抽取，生成语音的单帧语音保真度较高。但这种码本映射建立的转换函数时不连续的，容易导致语音内部频谱不连续。现已提出模糊矢量量化技术以及分段矢量量化技术等解决方案。

**8.5.2高斯混合模型法**

针对码本映射方法带来的离散性问题，在说话人识别领域中常用高斯混合模型来表征声学特征空间。高斯混合模型时传统方法中最主流的一种，它利用最小均方误差准则来确定转换函数，通过统计参数模型建立源说话人和目标说话人的映射关系，将源说话人的声音映射成目标说话人的声音。

与码本映射法相比，高斯混合模型有软聚类、增量学习和连续概率转换的特点。源声学特征和目标声学特征被看作联合高斯分布的观点被引入，通过使用概率论的条件期望思想获得转换函数，转换函数的参数皆可由联合高斯混合模型的参数估计算法得到。高斯混合模型转换的缺陷是：由于转换过程中常常进行求期望或加权平均的操作，转换后的特征往往过于平滑，听起来音质有所下滑。一种典型的补救方法是用全局方差进行正则化让转换后特征的方差符合训练数据特征方差的分布。

**8.5.3 深度神经网络法**

2010年以来，深度学习方法在智能语音领域得到广泛运用，在语音转换方面也出现了集中新的方法。除了在音质、相似性两方面超越传统方法以外，这些方法更重要的成就是打破了训练数据需要帧级对齐这个限制，其中有一些也打破了说话人身份固定这个限制。

比较典型的深层神经网络结构包括受限玻尔兹曼机-深层置信神经网络、长短时记忆递归神经网络、深度卷积神经网络等。由于深层神经网络具有较强的处理高维数据的能力，因此通常直接使用原始高维的谱包络特征训练模型，从而有助于提高转换语音的话音质量。与此同时，基于深度学习的自适应方法也被广泛应用于说话人转换，其利用少量新的发音人数据对已有语音合成模型进行快速自适应，通过迭代优化生成目标发音人的声音。

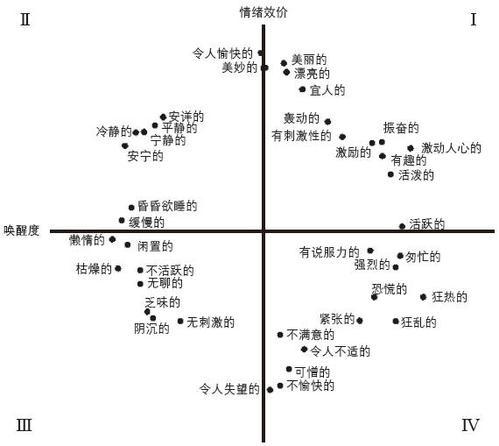
**8.6 情感语音**

语音不仅包含了语义信息，而且还携带了丰富的情感信息。自动语音情感识别则是计算机对人类情感感知和理解过程的模拟,它的任务就是从采集到的语音信号中提取表达情感的声学特征,并找出这些声学特征与人类情感的映射关系示；计算机的语音情感识别能力是计算机情感智能的重要组成部分,是实现自然人机交互界面的关键前提,具有很大的研究价值和应用价值。

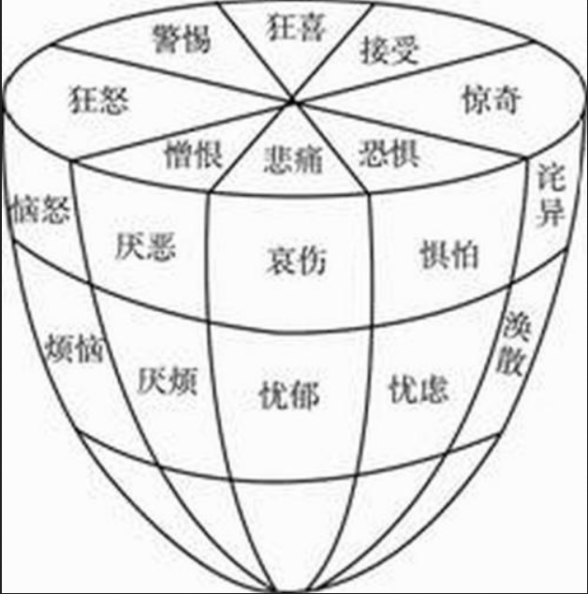
**8.6.1 情感描述**

离散情感模型将情感描述为离散的、形容词标签的形式,如高兴、愤怒等。丰富的语言标签描述了大量的情感状态，而用于研究的情感状态需要更具普遍性，因此人们定义了基本情感类别便于研究。其中,美国心理学家 Ekman 提出的 6 大基本情感(又称为 big six，即生气、厌恶、恐惧、高兴、悲伤和惊讶)在当今情感相关研究领域的使用较为广泛。

相对于离散情感模型，维度情感模型将情感状态描述为躲维度情感空间中的连续数值，也称为连续情感描述。将情感状态描述为多维情感空间中的点。这里的情感空间实际上是一个笛卡尔空间,空间的每一维对应着情感的一个心理学属性(例如,表示情感激烈程度的激活度属性以及表明情感正负面程度的效价属性)。理论上，该空间的情感描述能力能够涵盖所有的情感状态。换句话说，任意的、现实中存在的情感状态都可以在情感空间中找到相应的映射点，并且各维坐标值的数值大小反映了情感状态在相应维度上所表现出来的强弱程度。如图8.11(a)为情绪的二维模型，情感点同原点的距离体现了情感强度，相似的情感相互靠近；想法的情感则在二维空间中相距180°。图8.11(b)则是在二维空间中加入第三个维度强度后得到的三维情感空间模型。以强度、相似性和两级性划分情绪，越临近情绪性质越相似；距离越远差距越大。



(a) 二维模型



(b)三维模型

图8.11 情绪维度模型

**8.6.2 情感语音的声学特征**

情感语音中可以提取多种声学特征，用以反映人的情感行为的特点。用于语音情感识别的声学特征大致可归纳为韵律学特征、频谱特征和音质特征这三种类型。

韵律特征具有较强的情感辨别能力，它的情感区分能力已得到语音情感识别领域研究者们的广泛认可,使用非常普遍。其中最为常用的韵律特征有语速、基频、能量等。比如激动状态下语速就会比较快，喜、怒、惊等情感能量较大等等。但是韵律特征区分情感能力时十分有限的，例如愤怒、害怕、高兴和惊奇的基频特征具有相似的表现。

频谱特征被认为时声道形状变化和发声运动之间相关性的体现。语音中的情感内容对频谱能量在各个频谱区间的分布有着明显的影响.例如,表达高兴情感的语音在高频段表现出高能量,而表达悲伤的语音在同样的频段却表现出差别明显的低能量。频谱特征主要包括线性谱特征和倒谱特征。

音质特征是人们赋予语音的一种主观评价指标,用于衡量语音是否纯净、清晰、容易辨识等。对声音质量产生影响的声学表现有喘息、颤音、哽咽等,并且常常出现在说话者情绪激动、难以抑制的情形之下。用于衡量声质的声学特征一般有:共振峰频率及其带宽、频率微扰和振幅微扰、声门参数等。

**8.6.3 语音情感识别**

语音情感识别是让计算机能够通过语音信号识别说话者的感情状态，是情感计算的重要部分。情感计算的目的是通过赋予计算机识别、理解、表达和适应人的i情感的能力来建立和谐人机环境，使计算机具有更高的智能。

一般来说，语音情感识别系统由三部分组成：语音信号采集、语音情感特征提取和语音情感识别。语音信号采集模块通过语音传感器活得语音信号，并传递到语音情感特征提取模块；语音情感特征提取模块对语音信号中情感关联紧密的声学参数进行提取，最后送入情感识别模块完成情感判断。需要注意的使，语音情感识别离不开情感的描述和语音情感库的建立。当今语音情感识别系统所采用的识别算法可以分为如下两类：离散语音情感分类器，维度语音情感分类器。

离散语音情感分类器：它们一般被建模为标准的模式分类问题,即使用标准的模式分类器进行情感的识别。常用于语音情感识别领域的分类器,线性的有朴素贝叶斯，线性人工神经网络，线性支持向量机等;非线性有决策树，K-NN，高斯混合模型GMM ，隐马尔可夫模型HMM 以及稀疏表示分类器等。

维度语音情感分类器：该研究一般被建模为标准的回归预测问题,即使用回归预测算法对情感属性值进行估计,在当前的维度语音情感识别领域使用较多的预测算法有:线性回归,k-NN,ANN,支持向量回归等.其中,支持向量回归因为性能稳定、训练时间短等优点应用得最为广泛。

深度学习网络对语音情感识别也有所帮助，大致分为两类：一类是利用深度学习网络提出有效的情感特征，再送入分类器中进行识别；也有学者利用迁移学习的办法，在语音情感数据库上进行微调提取有效特征，获得良好效果。另一类研究者将分类器替换为深度神经网络进行识别，一些研究者将语音转化为语谱图送入卷积神经网络中，采用累死图像识别的处理方式为研究提供新思路。语音情感识别采用何种建模算法一直使研究者们非常关注的问题，但是在不同情感数据库上、不同的测试环境中，不同的识别算法各有优劣，不能一概而论。

**8.7课后练习**

1. 语音识别系统的组成有哪些？
2. 语音合成主要方法有哪些？每种方法的优缺点是什么？
3. 简单介绍语音转换在语音合成中的应用。
4. 语音情感识别系统由几部分组成？情感识别算法有哪几种？

[1]李德毅, 于剑, 工智能协会. 人工智能导论[M]. 中国科学技术出版社. 2018. 198-200