西安交通大学
XI'AN JIAOTONG UNIVERSITY

人工智能技术导论

第七章 语音处理

目录

西安交通大學

XI'AN JIAOTONG UNIVERSITY

7.1 语音的基本概念

语音的基本概念

语音是指人类通过发音器官发出来的、具有一定意义的、目的是用来进行**社会交际**的声音。语音是肺部呼出的气流通过在喉头至嘴唇的器官的各种作用而发出的。语音即语言的声音，是**语言符号系统的载体**，语言是音义结合的符号系统，语言的声音和语言的意义是紧密联系着的。

语言虽是一种声音，但又与一般的声音有着本质的区别。语音是人类发音器官发出的具有区别意义功能的声音，不能把语音看成纯粹的自然物质；语音是最直接地**记录思维活动的符号体系**，是语言交际工具的声音形式。

根据发音方式不同，将语音分为元音和辅音，辅音根据声带有无振动分为清辅音和浊辅音。人可感觉到**频率在20Hz—30kHz、强度为-5dB—130dB**的声音信号，在这个范围以外的音频分量是人耳听不到的。

语音的物理基础主要有**音高、音强、音长、音色**，这也是构成语音的四要素。

- **音高**：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，也称作“音质”。

语音的基本概念

语音经过采样以后，在计算机中以波形文件的方式进行存储，这种**波形文件反映语音在时域上的变化**。人们可以从语音的波形中判断语音音强（或振幅）、音长等参数的变化，但却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包括**傅立叶谱、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立叶谱，在此基础上根据人耳的听感特性，将语音信号在频域上划分成不同子带，进而可以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。

语音处理涉及许多学科，它以心理、语言和声学等为基础，以信息论、控制论和系统论等理论作为指导，通过应用信号处理、统计分析和模式识别等现代技术手段，发展成为新的学科。

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前景，而且正在逐渐改变人机交互的方式。

西安交通大學

XI'AN JIAOTONG UNIVERSITY

7.2 语音处理

语音处理主要有**语音合成、语音增强、语音转换、语音识别和情感语音**，本节主要介绍前三个，语音识别和情感语音在后两节专门介绍。

语音合成

语音合成也称文语转换，其主要功能是将任意的输入**文本转换成自然流畅的语音**输出。语音合成技术在银行、医院的信息播报系统和汽车导航系统、自动应答呼叫中心等都有广泛应用。

在语音合成技术中，主要分为**语言分析**部分和**声学系统**部分，也称为前端部分和后端部分。语言分析部分主要根据输入文字信息进行分析，生成对应的**语言学规格书**，想好该怎么读；声学系统部分主要根据提供的语音学规格书，**生成对应的音频**，实现发声的功能。

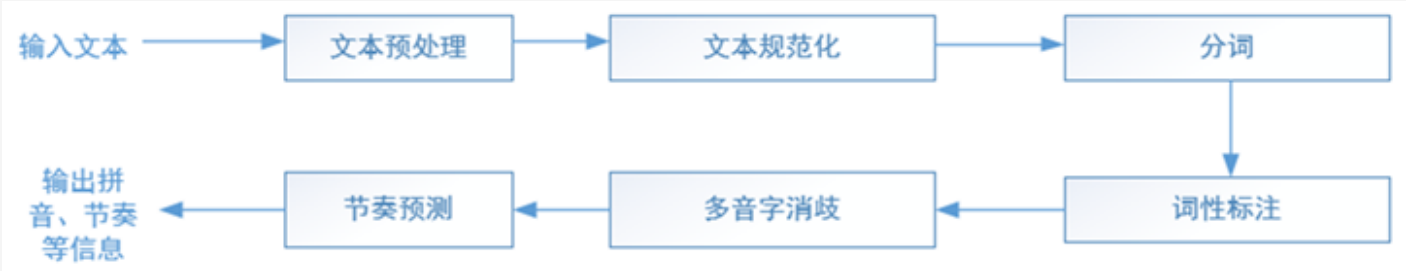
下图给出了一个基本的语音合成系统框图语音合成系统，包括**文本分析模块、韵律处理模块和声学处理模块**，其中文本分析模块可以视为系统的前端，而韵律处理模块和声学处理模块则视为系统的后端。

语音合成

◆ 文本分析模块

该模块是语音合成系统的前端，主要任务是对输人的任意文本进行分析，输出尽可能多的语言学信息（如拼音、节奏等），为后端的语音合成器提供必要的信息。对简单系统而言，只需提供拼音信息足够；而对高自然度的合成系统，文本分析需要给出更详尽的语言学和语音学信息（实际上属于自然语言理解范畴）。

对于汉语语音合成系统，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的表达。
- 自动分词是将待合成的整句以词为单位划分为单元序列，以便后续考虑词性标注、韵律边界标注等。
- 词性标注为每个词汇分配词性标签，如名词、动词、形容词等，因为词性可能影响字或词发音方式。
- 字音转换的任务是将待合成的文字序列转换为对应的拼音序列，即告诉后端合成器应该读什么音。由于汉语中存在多音字问题，所以字音转换的一个关键问题就是处理多音字的消歧问题。

语音合成

◆ 韵律处理模块

直观来讲，韵律即是实际语流中的抑扬顿挫和轻重缓急，如重音的位置分布及其等级差异，韵律边界的位置分布及其等级差异，语调的基本骨架及其跟声调、节奏和重音的关系等。韵律表现是一个复杂现象，对韵律的研究涉及语音学、语言学、声学、心理学、物理学等多个领域。但是，作为语音合成系统中承上启下的模块，韵律模块实际上是语音合成系统的核心部分，极大地影响着最终合成语音的自然度。

说话人通过在语音不同位置进行停顿来准确表达语义和意图。停顿将中文语音文本分割为不同的韵律成分，分别是韵律词（prosodic word, PW），韵律短语（prosodic phrase, PPH）和语调短语（intonational phrase, IPH）。韵律结构预测对语音合成的自然度和可懂度有重要作用，（从听者的角度来看）与韵律相关的语音参数包括基频、时长、停顿和能量等。

现有方法把韵律结构预测视为一个序列到序列的分类任务，即模型需要预测每个字符后是否存在一个韵律停顿，如使用人工设计的特征、词向量、BERT、语法信息作为输入特征，还尝试使用BLSTM-CRF模型、最大熵模型、基于自注意力机制的模型、基于多任务学习进行韵律结构预测。同时一些研究揭示在篇章语音中存在高于句子级别的韵律模式，即来自上下文其他语句的文本信息能够提升韵律结构预测的精度。

语音合成

◆ 声学处理模块

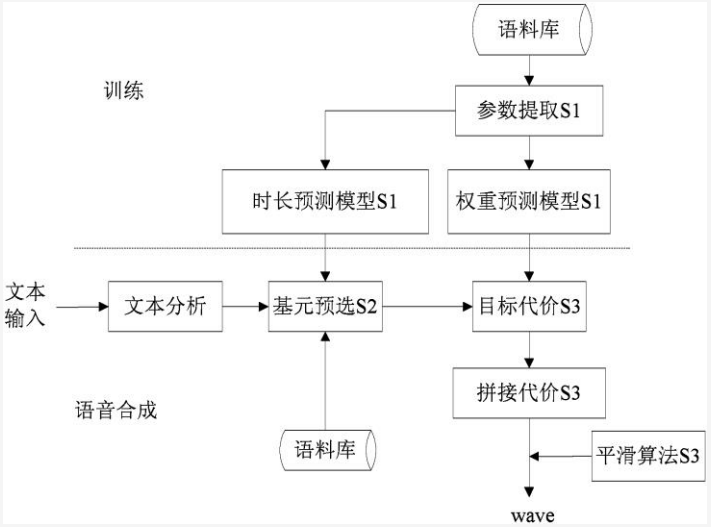
声学处理模块根据文本分析模块和韵律处理模块提供的信息来生成自然语音波形。语音合成系统的合成阶段可以简单概括为两种方法：

1）基于时域波形的拼接合成方法：

根据韵律处理模块提供的基频、时长、能量和节奏等信息，在预先录制并标注好的语音库中挑选合适语音基元进行适度调整，然后通过拼接算法生成语音波形。

基元是指用于语音拼接时的基本单元，可以是音节或者音素等。受限于计算机存储能力与计算能力，早期的拼接合成方法的基元库都很小，同时为了提高存储效率，往往需要将基元参数化表示；此外，由于拼接算法本身性能的限制，常导致合成语音不连续、自然度很低。

随着计算机运算和存储能力的提升，**实现基于大语料库的基元拼接合成系统成为可能**。在这种方法中，基元库由以前的几MB扩大到几百MB甚至是几GB。由于大语料库具有较高的上下文覆盖率，使挑选出来的基元几乎不需要做任何调整就可用于拼接合成。因此，相比于传统的参数合成方法，该方法合成语音在音质和自然度上都有了极大的提高，而基于大语料库的单元拼接系统也得到了十分广泛的应用



拼接合成方法依旧存在着些不足——稳定性仍然不够，拼接点不连续的情况还是可能发生；难以改变发音特征，只能合成该建库说话人的语音。由于基于波形拼接的语音合成方法存在着一些固有的缺陷，限制了其在多样化语音合成方面的应用，因此，基于参数合成的可训练语音合成方法被提出。

语音合成

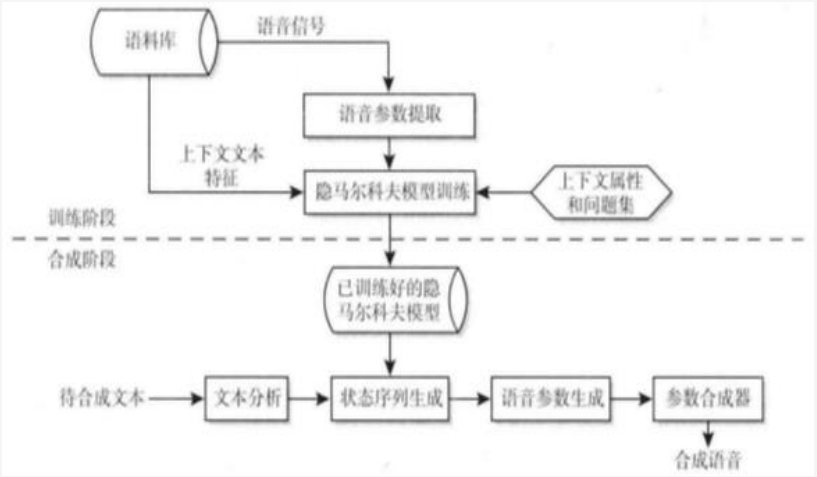
◆ 声学处理模块

2）基于语音参数的合成方法

该方法根据一定的语音数据进行训练并快速构建合成系统，而且对于不同发音人、不同发音风格甚至不同语种的依赖性非常小，非常符合多样化语音合成方面的需求。其中最成功的是基于隐马尔可夫模型的可训练语音合成方法，主要分为训练阶段和合成阶段两个阶段。

在隐马尔可夫模型训练前，首先要对一些参数进行配置，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。

模型训练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**问题集进行设计**，即根据先验知识选择一些对语音参数有一定影响的上下文属性并设计相应的问题集，如前后调、前后声韵母等。需要注意的是，这部分工作是与语种相关的。除此之外，**整个流程基本上与语言种类无关**。



随着深度学习的研究进展，深度神经网络也被引入统计参数语音合成中，以代替该方法中的隐马尔科夫模型，可直接通过一个深层神经网络来预测声学参数，克服了隐马尔可夫模型训练中决策树聚类环节中模型精度降低的缺陷，进一步增强了合成语音的质量。由于基于深度神经网络的语音合成方法体现了比较高的性能，目前已成为参数语音合成的主要方法。

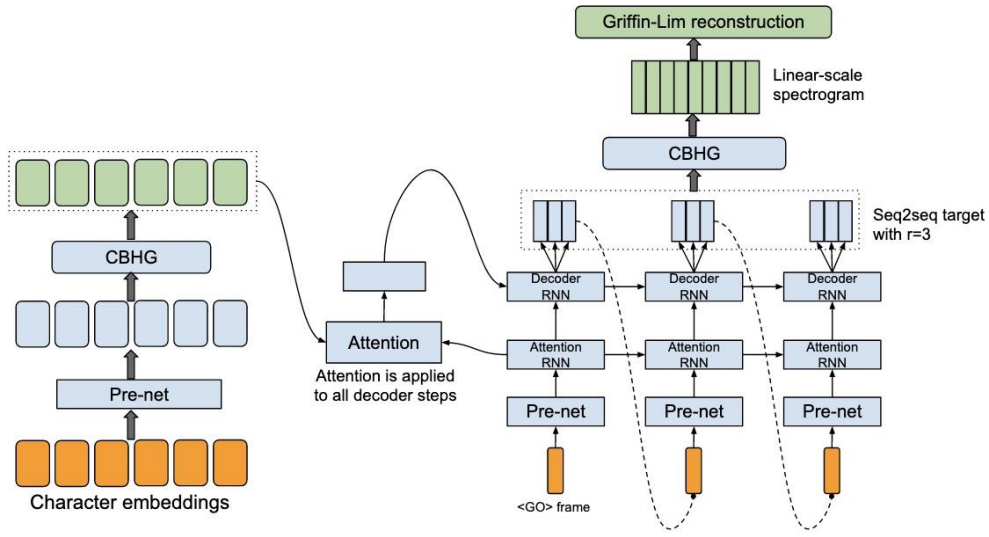
语音合成

◆ 声学处理模块

□ 基于端到端的语音合成方法

2017年初，Google 提出了一种端到端的语音合成系统——Tacotron。所谓“端到端”就是直接从字符文本合成语音，打破各个传统组件之间壁垒，可从<文本，声谱>配对的数据集上，完全随机从头开始训练。

该模型主要是基于带有注意力机制的编码-解码模型。其中，编码器是一个以字符或者音素为输人的神经网络模型；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。



这种方法的自然度和表现力已经能够媲美于人类说活水平，并不需要多阶段建模过程，已经成为当下热点和未来发展趋势。

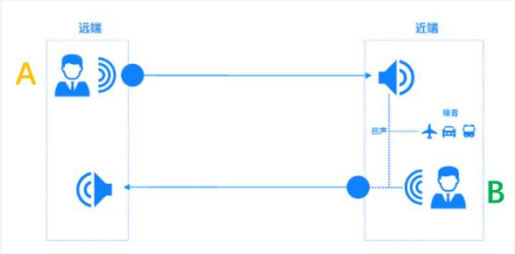
语音增强

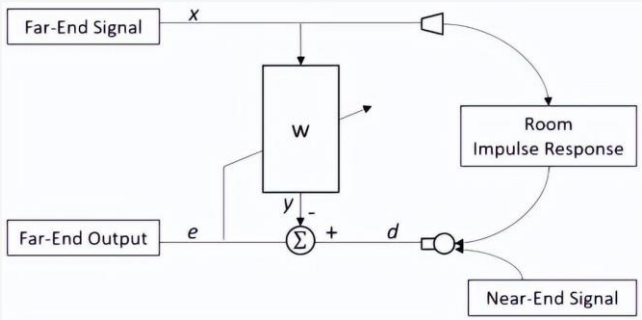
语音增强是指当语音信号被各种各样的干扰源淹没后，从混叠信号中提取出有用的语音信号，抑制、降低各种干扰的技术，主要包括回声消除、混响抑制、语音降噪等关键技术。

◆ 回声消除

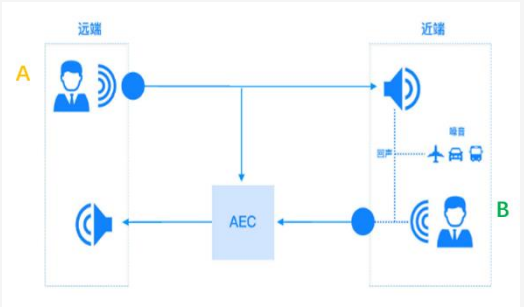
回声干扰是指远端扬声器播放的声音经过空气或其他介质传播到近端的麦克风形成的干扰。回声消除最早应用于语音通信中，终端接收的语音信号通过扬声器播放后，声音传输到麦克风形成回声干扰。

远端讲话者（A）的声音信号再传输给近端（B）后，在近端设备的扬声器播放出来，经过一系列声学反射，被近端设备的麦克风拾取，又传输给远端（A）的现象。声学回声将导致远端讲话者A在很短时间内，又听到了自己刚才的讲话声音。声学回声的产生过程如右图所示。使用 AEC 技术后，两端声音传输过程改变为如右下图所示，进而从底层保证会议场景下声音的干净度。





数学化表示如左图，远端信号x, 从听筒或喇叭播出，并经过空间传播，被麦克风接收，近端说话信号也进入麦克风，这样麦克风接收到的就是两个信号的叠加，即***d***。**自适应滤波器*w***对x进行处理获得***y***（回声信号），***d***和***y***的差值作为**误差*e***，传递给自适应滤波器，进行滤波器系数迭代更新。



引文链接：https://zhuanlan.zhihu.com/p/632443142

语音增强

◆ 混响抑制

混响干扰是指声音在房间传输过程中，会经过墙壁或其他障碍物的反射后通过不同路径到达麦克风形成的干扰源。房间大小、声源和麦克风的位置、室内障碍物、混响时间等因素均影响混响语音的生成。

在声源停止后，混响声可以在一段时间内听到，因为它变得越来越柔和。**混响声相对于直达声的振幅被称为饱满度**。清晰度与饱满度相反，是通过降低混响声的振幅来实现的。“饱满”通常意味较长的混响时间，而“清晰”则意味着短的混响时间。

在声源停止后，绝对强度下降106倍所需时间，或说，**强度下降60分贝所需时间，被定义为混响时间**（RT，有时被称为RT60）。混响时间过短，则声音发干、枯燥无味、不亲切自然；混响时间过长，声音含混不清；混响时间合适时声音圆润动听。

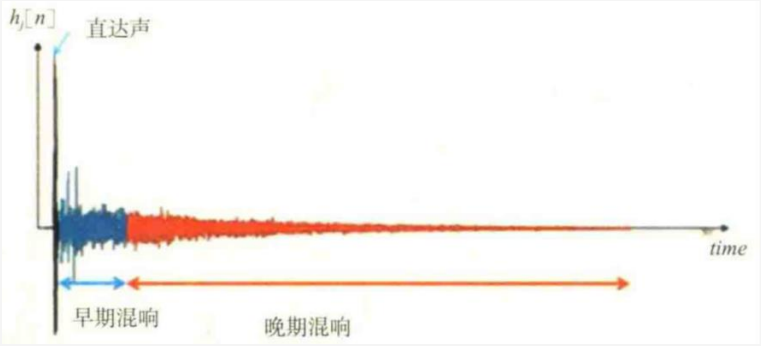
大多数房间的混响时间在200-100毫秒。右图为一个典型房间脉冲响应，蓝色部分为早期混响，橙色部分为晚期混响。

在语音去混响任务中，重多地关注于对晚期混响的抑制。目前，去混响的方法有三种：

1）利用声学系统（或房间）的数学模型消除混响，并在估计房间声学模型参数后，形成对原始信号的估计。

2）通过将混响作为一种（卷积）噪声处理，并进行专门适应混响的去噪处理，抑制混响。

3）直接使用深度神经网络机器等学习方法或备选的多通道线性滤波器，从麦克风信号中估计原始去混响信号。



语音增强

◆ 语音降噪

语音降噪是指通过技术手段减少或消除背景噪声，以提升语音信号的清晰度和可理解性。这一过程不仅涉及对噪声的抑制，还包括确保语音的自然和真实感。

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。

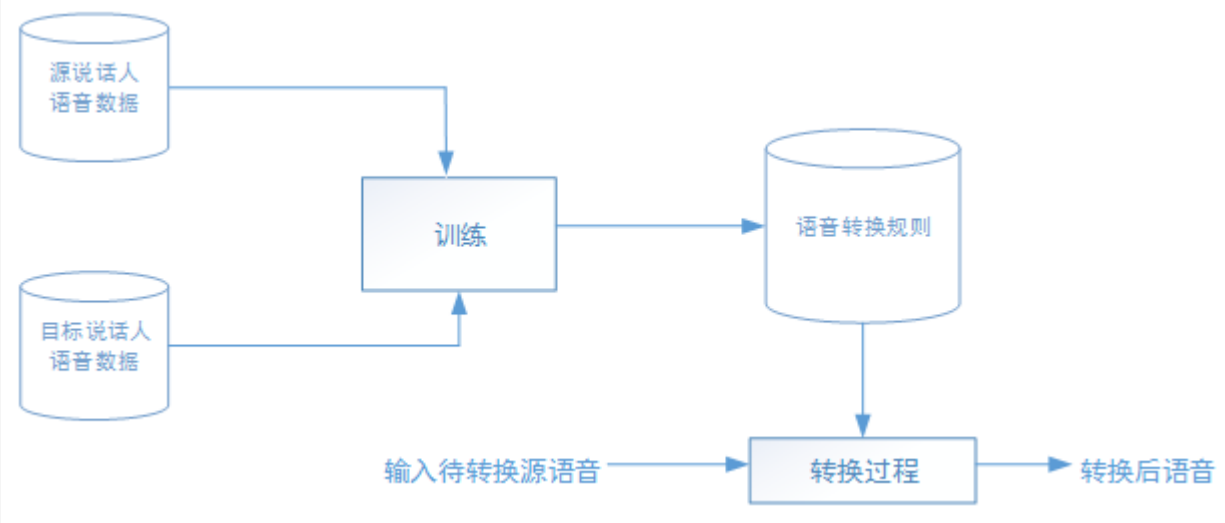
多通道语音降噪的目的是融合多个通道的信息，抑制非目标方向的干扰源，增强目标方向的声音。通常受限于麦克风阵列的结构，比较典型的阵列结构包括线阵和环阵。麦克风阵列的选型与具体的应用场最相关。随着麦克风个数的增多，噪声抑制能力会更强，但算法复杂度和硬件功耗也会相应增加。

语音转换

语音转换是通过语音处理手段改变语音中的说话人个性信息，使改变后的语音听起来像是由另外一个说话人发出的。首先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。

实现完整的语音转换包括离线训练和在线转换两个阶段，在训练阶段，首先提取源说话人和目标说话人的个性特征参数，然后根据某种匹配规则建立源说话人和目标说话人之间的匹配函数；在转换阶段，利用训练阶段获得的匹配函数对源说话人的个性特征参数进行转换，最后利用转换后的特征参数合成出接近目标说话人的语音。

基本框图如下所示：



语音转换

- **码本映射法。**

最早应用于语音转换的方法，且一直到现在仍有很多研究人员使用这种转换算法。在这种方法中，源码本和目标码本的单元一对应，通过从原始语音片段中抽取关键的语音顿作为码本，建立起源说话人和目标说话人参数空间的关系。其优点在于码本从原始语音片段中抽取，生成语音的单顿语音保真度较高。但这种码本映射建立的转换函数是不连续的，容易导致语音内部频谱不连续。

- **高斯混合模型法。**

针对码本映射带来的离散性问题，常用高斯混合模型来表征声学特征空间。这种方法使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。与码本映射方法相比，高斯混合模型有软聚类、增量学习和连续概率转换的特点。在高斯混合模型算法中，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人，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，此时高斯混合模型映射方法成为频谱转换研究的主流映射算法。高斯混合模型转换方法的缺点是会给转换特征带来过平滑的问题，导致转换语音的音质下降。

- **深度神经网络法。**

通过深层神经网络模型建立源说话人和目标说话人之间的映射关系，实现说话人个性信息的转换，解决高斯混合模型方法中的过平滑问题。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。此外，我们还可以通过语音转换技术去除说话人的个性信息，将说话人语音编程机器声或沙哑声，保护说话人的隐私。

西安交通大学

XI'AN JIAOTONG UNIVERSITY

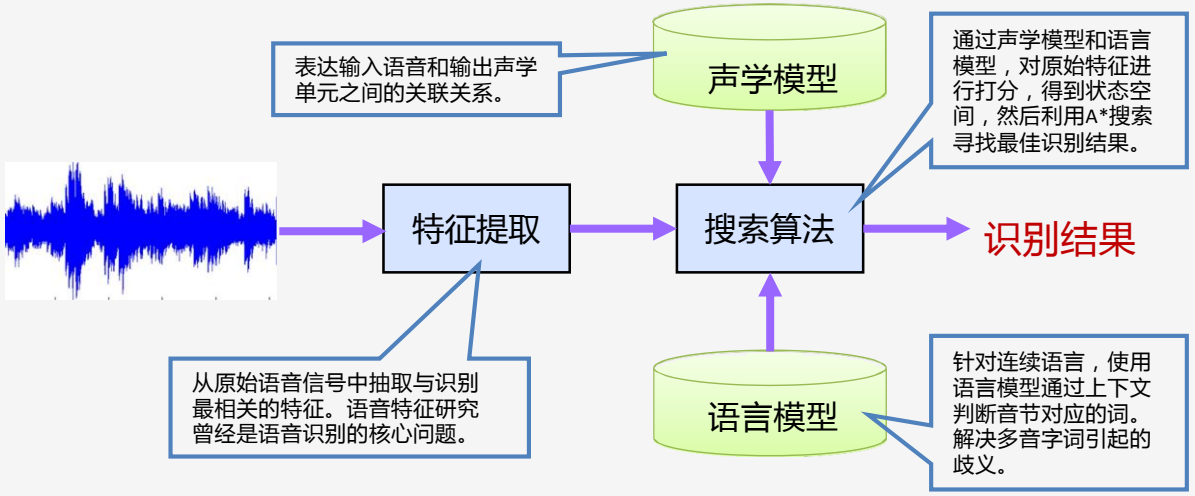
7.3 语音识别

语音识别

语音识别是指将语音自动转换为文字的过程。在实际应用中，语音识别通常与自然语言理解、自然语言生成及语音合成等技术相结合，提供个基于语音的自然流畅的人机交互系统。

- 语音识别技术的发展历史可以追溯到20世纪50年代初期。1952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，其中基于混合高斯的隐马尔科夫模型（GMM-HMM）最为成功。
- 21世纪的前十年是语音识别技术的打磨期。尽管没有划时代的技术革新，但大量真实数据的运用极大地提高了系统的实用性。另一方面，机器学习技术被广泛应用，以区分性训练为代表的新方法将GMM-HMM系统的性能发挥到了极致。这一阶段，语音识别开始走向实用化。
- 2011年至今是深度学习阶段，在实际应用场景下，基于大规模数据训练的深度神经网络（DNN）在某些领域甚至已经超过人类的识别水平。今天，DNN已经成为语音识别的主流模型，统治业界30年的GMM-HMM模型已然退出历史舞台。

语音识别系统主要包括四个部分：特征提取、声学模型、语言模型和解码搜索。



语音识别

特征提取

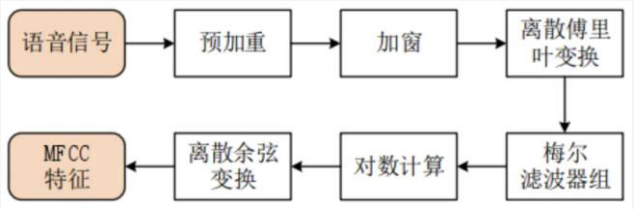
语音特征抽取即是在原始语音信号中提取出与语音识别最相关的信息，滤除其他无关信息，比较常用的声学特征是梅尔频率倒谱系数（Mel-scale Frequency Cepstral Coefficients，简称MFCC）。

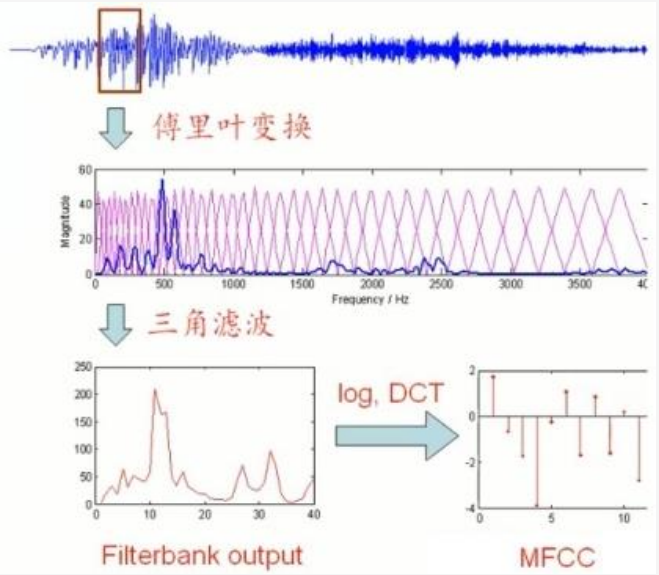
人耳对不同频率的声波有不同的听觉敏感度，两个响度不等声音作用于人耳时，响度较高的频率成分存在会影响到对响度较低的频率成分的感受，使其变得不易察觉，这种现象称为掩蔽效应。心理物理学研究表明，人类对语音信号频率内容的感知遵循一种主观上定义的非线性尺度，该非线性标度可被称为“Mel”标度。

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，它与频率的关系可用下式近似表示：

$$\mathrm{Mel}(f) = 2595 \times \lg(1 + \frac{f}{700})$$

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的假设和限制，又利用了听觉模型的研究成果，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。





引文链接：https://blog.csdn.net/qq_36002089/article/details/120014722
https://download.csdn.net/blog/column/10767007/115499309

语音识别

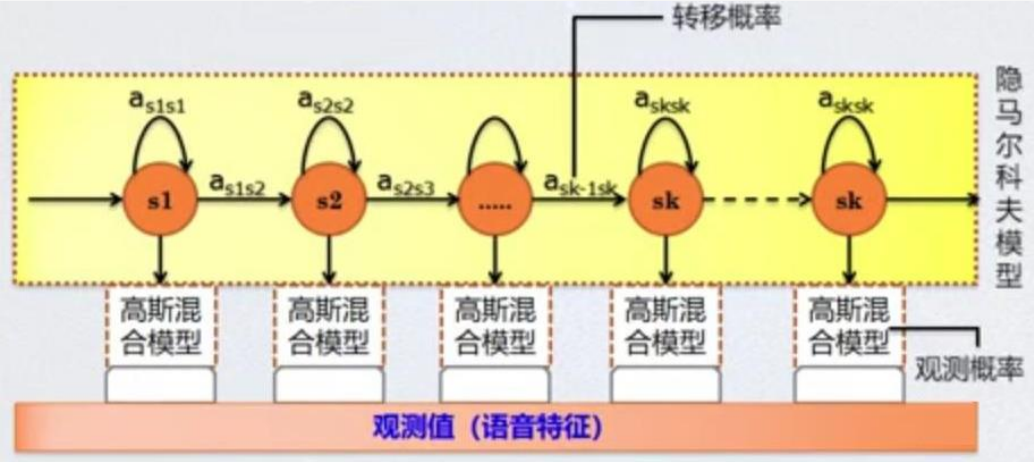
声学模型

声学模型承载着声学特征与建模单元之间的映射关系。在训练声学模型之前需要选取建模单元，建模单元可以是音素、音节、词语等，其单元粒度依次增加。目前大多数声学模型一种般采用音素作为建模单元，语音中存在协同发音的现象，即音素是上下文相关的，故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，即基于高斯混合模型-隐马尔科夫模型的模型和基于深度神经网络-隐马尔科夫模型的模型。

1）基于高斯混合模型-隐马尔科夫模型的声学模型

隐马尔科夫模型的参数主要包括状态间的转移概率以及每个状态的概率密度函数，也叫出现概率，一般用高斯混合模型表示。就基于高斯混合模型-隐马尔科夫模型的声学模型而言，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；对于中等和大词汇量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。

该声学模型的框架图如右图所示，高斯混合模型用来估计观察特征（语音特征）的观测概率，而隐马尔科夫模型则被用于描述语音信号的动态变化（即状态间的转移概率）。



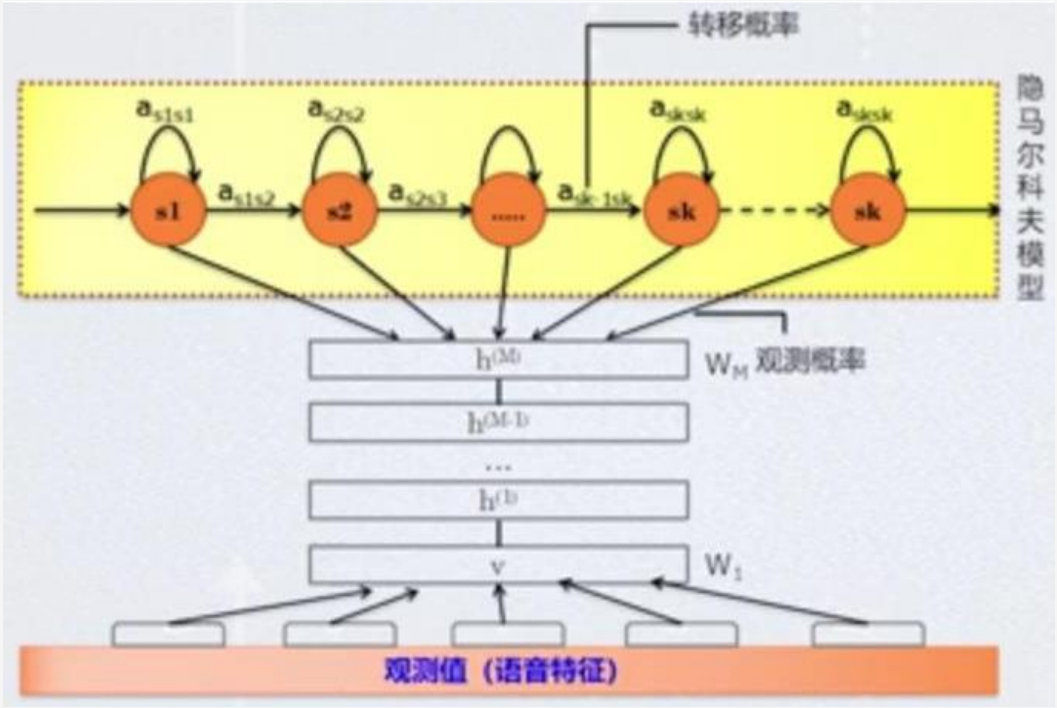
声学模型

2）基于深度神经网络-隐马尔科夫模型的声学模型

该模型是指用深度神经网络模型替换上述模型的高斯混合模型，深度神经网络模型可以是深度循环神经网络和深度卷积网络等。该模型的建模单元为聚类后的三音素状态。

与基于高斯混合模型相比，这种基于深度神经网络的声学模型具有两方面的优势：一是深度神经网络能利用语音特征的上下文信息；二是深度神经网络能学习非线性的更高层次特征表达。

故此，基于深度神经网络-隐马尔科夫模型的声学模型的性能显著超越基于高斯混合模型-隐马尔科夫模型的声学模型，已成为目前主流的声学建模技术。



语言模型

语音模型是根据语言客观事实而进行的语音抽象数学建模。语言模型也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 =“今天天气怎么样”，该句子很常见，通过语言模型可以算出其发生的概率P（今天天气怎么样）=0.80000

例2：令句子S =“材教人工智能”，该句子是病句，不常见，通过语言模型可算出其发生的概率 P（材教人工智能）=0.00001

语言模型是用来约束单词搜索的，它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。大部分的语言模型都是使用n-gram模型，它包含了单词序列的统计和有限状态模型，它通过有限状态机来定义语音序列。

对于一个句子 $w_{1:N}=\{w_1, w_2, w_3, ..., w_N\}$, w_i 代表词，$i=1,2,...,N$。在真实的语言模型中，w_i 也可以是token等形式，例如词语。句子 $w_{1:N}$ 出现的概率的计算公式如下所示：

$$P_{n-grams}(w_{1:N}) = \prod_{i=n}^{N} \frac{C(w_{i-n+1:i})}{C(w_{i-n+1:i-1})}$$

其中：

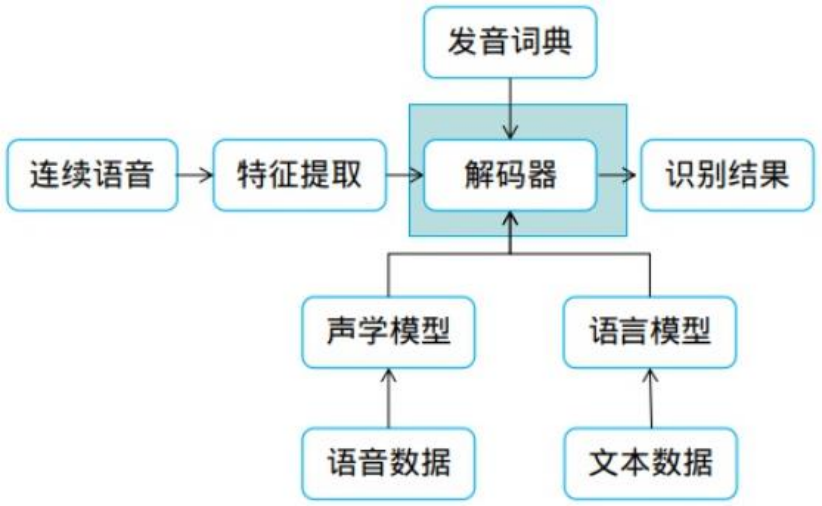
- $P_{n-grams}(w_{1:N})$ 是句子 $w_{1:N}$ 在n-grams模型下的概率。
- $C(w_{i-n+1:i})$ 是在语料库中，连续n个词 $w_{i-n+1:i}$ 出现的次数。
- $C(w_{i-n+1:i-1})$ 是在语料库中，连续n-1个词 $w_{i-n+1:i-1}$ 出现的次数。

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的程度。如果我们对于某件事情知道得越多，那么困惑度越小，因此构建语言模型时，目标就是寻找困惑度较小的模型，使其尽量逼近真实语言的分布。

解码搜索

解码搜索的主要任务是在由声学模型、发音词典和语言模型构成的搜索空间中寻找最佳路径。

解码时需用到声学得分和语言得分，声学得分由声学模型计算得到，语言得分由语言模型计算得到。其中，每处理一帧特征都会用到声学得分，但是语言得分只有在解码到词级别才会涉及，一个词一般覆盖多帧语音特征。故此，解码时声学得分和语言得分存在较大数值差异。为了避免这种差异，解码时将引入一个参数对语言得分进行平滑，从而使两种得分具有相同的尺度。



构建解码空间的方法可以概括为两类：静态的解码和动态的解码。静态的解码需要预先将整个静态网络加载到内存中，因此需要占用较大的内存。动态的解码是指在解码过程中动态地构建和销毁解码网络，这种构建搜索空间的方式能减小网络所占的内存，但是基于动态的解码速度比静态慢。通常在实际应用中，需要权衡解码速度和解码空间来选择构建解码空间的方法。解码所用的搜索算法大概分成两类：采用**时间同步**的方法，如维特比算法等；**时间异步**的方法，如A*算法等。

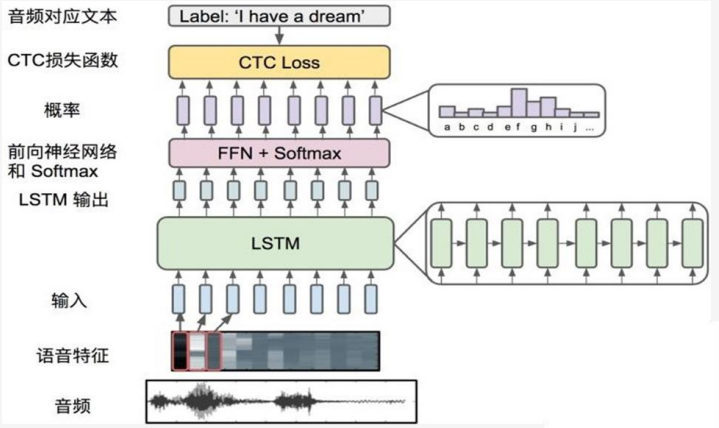
语音识别

基于端到端的语音识别方法

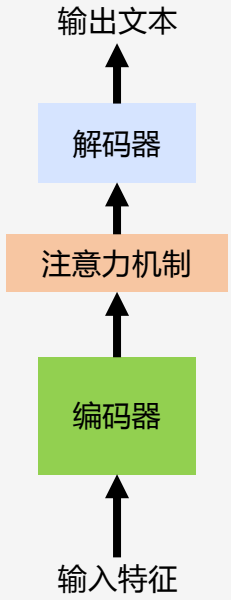
上述混合声学模型存在两点不足：一是神经网络模型的性能受限于高斯混合模型-隐马尔科夫模型的精度；二是训练过程过于繁复。为了解决这些不足，研究人员提出了端到端的语音识别方法，一类是基于联结时序分类的端到端声学建模方法；另一类是基于注意力机制的端到端语音识别方法。前者只是实现声学建模的端到端，后者实现了真正意义上的端到端语音识别。

① **基于联结时序分类的端到端方法**是在声学模型训练过程中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，因此不需要高斯混合模型-隐马尔科夫模型生成强制对齐信息，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，极大地简化了声学模型训练的过程。

但是语言模型还需要单独训练，从而构建解码的搜索空间。而循环神经网络具有强大的序列建模能力，所以联结时序分类损失函数一般与长短时记忆模型结合使用，当然也可和卷积神经网络的模型一起训练。



② **基于注意力机制的端到端语音识别方法**实现真正的端到端。传统的语音识别系统中声学模型和语言模型是独立训练的，但是该方法将声学模型、发音词典和语言模型联合为一个模型进行训练。端到端的模型是基于循环神经网路的编码-解码结构。



编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的有用信息，而解码器则将该定长序列扩展成输出单元序列。尽管这种模型取得了不错的性能，但其性能远不如混合声学模型。谷歌提出了一种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。

西安交通大學

XI'AN JIAOTONG UNIVERSITY

7.4 情感语音

情感语音

语音作为人们交流的主要方式，不仅包含语义信息，而且还携带有丰富的情感信息。**语音信号是语言的声音表现形式，情感是说话人所处环境和心理状态的反映**，同样一句话，如果说话人的情感和语气不同，听者的感知也有可能会不同，分析和处理语音信号中的情感信息对判断说话人的喜怒哀乐具有重要意义。

情感描述

研究语音信号的情感，首先要根据某些特性标准对情感做一个有效合理的分类，然后在不同类别的基础上研究特征参数的特性。

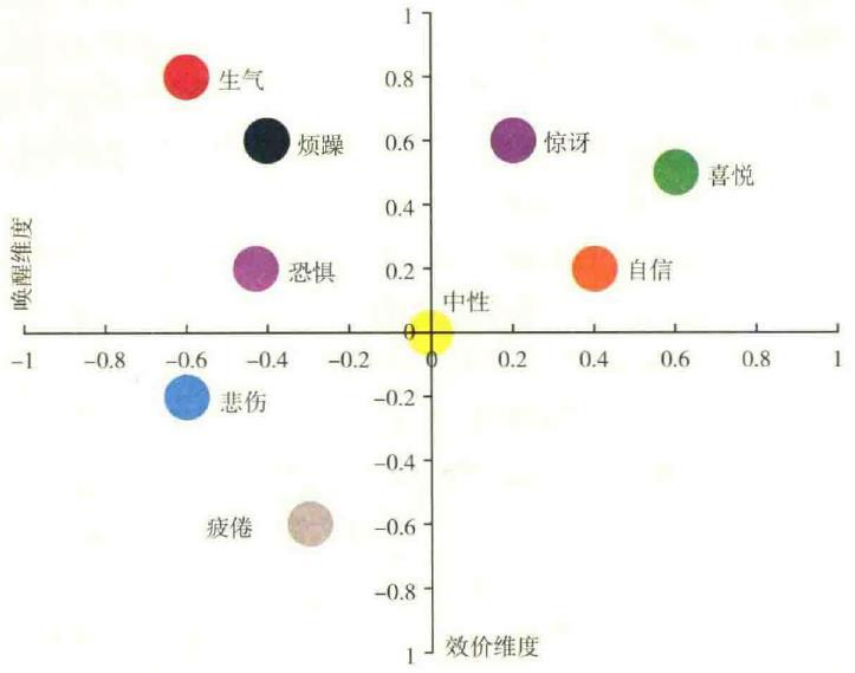
目前，主要从**离散情感和维度情感**两个方面来描述情感状态。离散情感模型将情感描述为离散的、形容词标签的形式，如高兴、愤怒等。美国心理学家Ekman提出的六大基本情感（生气、厌恶、恐惧、高兴、悲伤和惊讶）在当今情感相关研究领域的使用较为广泛。

相对于离散情感模型，维度情感模型将情感状态描述为多维情感空间中的连续数值，也称作连续情感描述。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**（如表示情感激烈程度的激活度属性、表明情感正负面程度的愉悦度属性）。

情感描述

Russel等人在激活愉悦空间上用一个情感轮对情感进行分类，左下图所是情绪的二维模型。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；相反的情感则在二维空间中相距180度。

当在这个二维空间中加入强度作为第三个维度后，可以得到右下图的三维的情感空间模型。以强度、相似性和两极性划分情绪，模型上方的圆形结构划分为八种基本情绪：狂喜、警惕、悲痛、惊奇、狂怒、恐惧、接受和憎恨。越邻近的情绪性质上越相似，距离越远，差异越大，互为对顶角的两个扇形中的情绪相互对立。



声学特征

情感语音中可以提取多种声学特征，用以反映说话人的情感行为的特点。情感特征的优劣对情感处理效果的好坏有重要影响。语音声学情感特征主要分为三类：韵律特征、音质特征以及频谱特征。

情感状态与一些语音参数的关系如下表所示。此外，基于这三类语音特征的不同语段长度的统计特征是目前使用最为普遍的特征参数之一，如特征的平均值、变化率、变化范围等。

特征参数	悲伤	愤怒	高兴	惊奇	平静
平均持续时间	很长	很短	较长	较短	一般
发音速率	很慢	很快	较慢	较快	一般
幅度均值	较小	较大	较大	很大	一般
幅度范围	较小	很大	很大	很大	一般
基频平均值	一般	较大	较大	很大	一般
基频动态值	一般	较大	很大	很大	一般
基频变化率	较慢	很快	较快	很快	一般

引文链接：https://zhuanlan.zhihu.com/p/496818604?utm_id=0

声学特征

◆ 韵律特征

该特征并不影响对语音语义信息的识别，但决定着语音流畅度、自然度和清晰度。最常用的韵律特征有：

- **时长相关特征：**如语速、短时平均过零率等
- **基频相关特征：**如基因频率及其均值、变化范围、变化率、均方差等
- **能量相关特征：**短时平均能量、短时能量变化率、短时平均振幅等

➢ **语速：**文本中元音持续时间与元音数目的比值。情绪高涨（高兴、愤怒）时语速快，情绪消沉（伤心、难过）时语速较慢。

$$t_{avg} = \frac{1}{m}\sum_{i=1}^{m} t_i$$

其中，***m***表示语音中所包含的元音数目，***i***代表第***i***个元音，t_i代表第***i***个元音的持续时间。

➢ **短时平均能量：**短时平均能量与声音震动的幅值相关，描述的是语音信号的能量值，且发生在相对短的时间内。在一般情况下，如果讲话人讲话的声音大，则消耗的能量就比较大；如果讲话人的声音较小声，代表消耗的能量比较小。对应到不同情感中时，一般在生气惊讶等发出的音量很大，即语音的能量变大，在伤心失落或平静时，语音的音量变低，即语音的能量变小。

$$\begin{aligned} E_m &= \sum_{n=-\infty}^{\infty}[x(n)w(m-n)]^2 \\ &= \sum_{n=m-N+1}^{m}[x(n)w(m-n)]^2 \end{aligned}$$

其中，***Em***代表第***m***帧语音信号的短时能量值，***w(m)***表示窗函数，窗长为***N***，***x(n)***代表语音信号。

➢ **短时平均过零率：**它代表的是每一个分帧内语音信号幅度值为零的次数。语音信号的短时平均过零率特征一定程度地能够描述信号的频率谱特性，因此能够大致估算谱的特性。

$$\begin{aligned} Z_m &= \sum_{n=-\infty}^{\infty} |\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]| w(m-n) \\ &= |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]| * w(m) \end{aligned}$$

$$w(m) = \begin{cases} \frac{1}{2N}, & 0 \le m \le N-1 \\ 0, & \text{其他} \end{cases}$$

$$\operatorname{sgn}[x(n)] = \begin{cases} 1, & x(n) \ge 0 \\ -1, & x(n) < 0 \end{cases}$$

声学特征

◆ 音质特征

该特征是语音的一种主观评价指标，描述了声门激励信号的性质，包括发声者语态、喘息、颤音及哽咽，用来衡量语音纯净度、清晰度和辨识度。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。声音质量的变化被听辨者们一致认定为与语音情感的表达有着密切的关系。通过对声音质量的评价，可获得说话人的生理、心理信息并对其情感状态进行区分。用于衡量声音质量的声学特征一般有：共振峰频率、带宽、频率扰动、振幅扰动、谐波噪声比、闪光及声门参数等

◆ 频谱特征

该特征体现了声道形状变化与发声运动间的相关性。谱特征参数反映信号在频域的特性，不同情感在各个频谱间的能量是有差异的（如表达欢快的语音在高频区间能量较高，表达哀愁的语音在同样的频段能量较低）。基于谱的相关特征主要分为：

- **线性频谱特征：**线性预测系数（Linear Prediction Coffcients，LPC）、对数频率功率系数（Log Frequency Power Coefficients，LFPC）及单边自相关线性预测系数（One—sided Auto correlation Linear Predictor Coefficient，OSALPC）等；
- **倒谱特征：**线性预测倒谱系数（Linear Prediction Cepstrum Cofficiems，LPCC）、单边自相关线性预测倒谱系数（One—sided Autocorrelation Linear Predictor Cepstral ·—based Coefficient，OSALPCC）以及梅尔频率倒谱系数（MFCC）等

语音情感识别

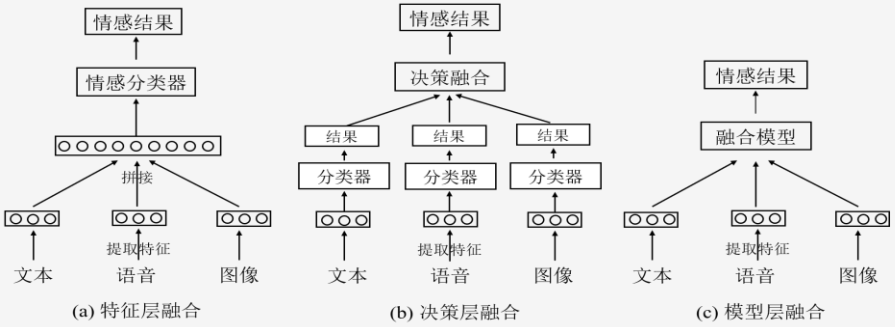
语音情感识别是让计算机能够通过语音信号识别说话者的情感状态，是情感计算的重要组成部分，是情感语音处理的主要内容之一。情感计算的目的是通过赋予计算机识别、理解、表达和适应人的情感的能力来建立和谐人机环境，并使计算机具有更高的、全面的智能。情感语音利用语音信息进行情感计算。

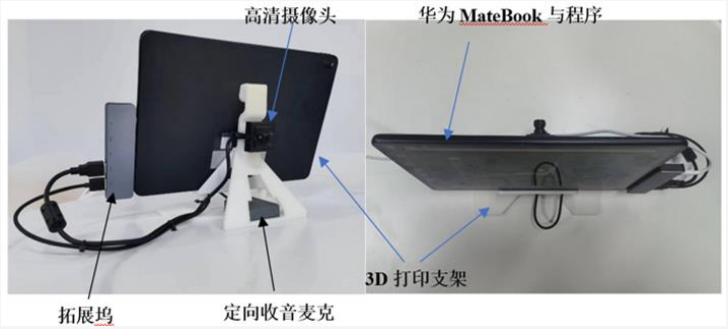
语音情感识别本质是一个典型的模式分类问题，因此模式识别领域中的诸多算法都可用于语音情感识别研究。近年来语音情感识别领域有三种主流思路：一是使用手工提取的特征作为输入，用传统机器学习算法进行分类或是使用人工神经网络进行分类；二是以原始信号直接作为输入，利用人工神经网络内部的结构与深度自行学习特征，进行分类。 三是基于混合模型的方法，这种方法结合了特征提取和深度学习的方法。首先通过特征提取获得初步的特征表示，然后使用深度学习模型对这些特征进行进一步的处理和分类。

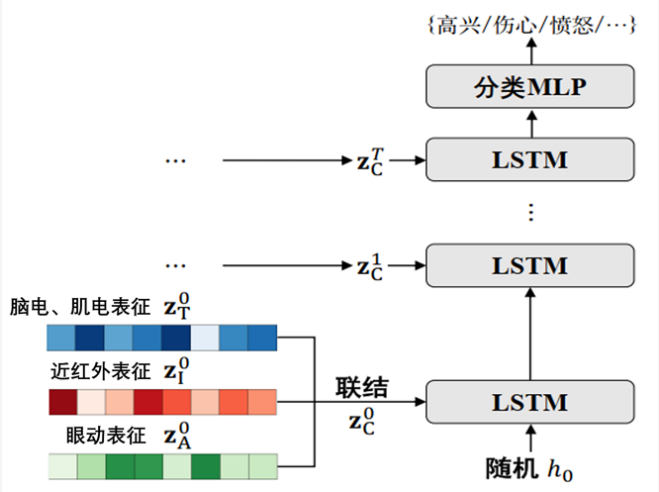
然而，由于人们表达情感方式的多样性以及表达情感不依赖固定范式，情感的表达形式随着场景及个人性格的不同而不同，因此仅仅依赖某一种模态数据进行情感计算所获得的结果并不全面，这是单模态情感分析的固有缺陷。多模态情感分析通过提取文本、语音、人脸表情以及生理信号等的多种模态数据的情感特征，完成情感的分类与预测，既要保证充分挖掘各模态内部的情感语义，又要保证跨模态情感信息的良好交互，实现不同模态情感语义的信息互补。

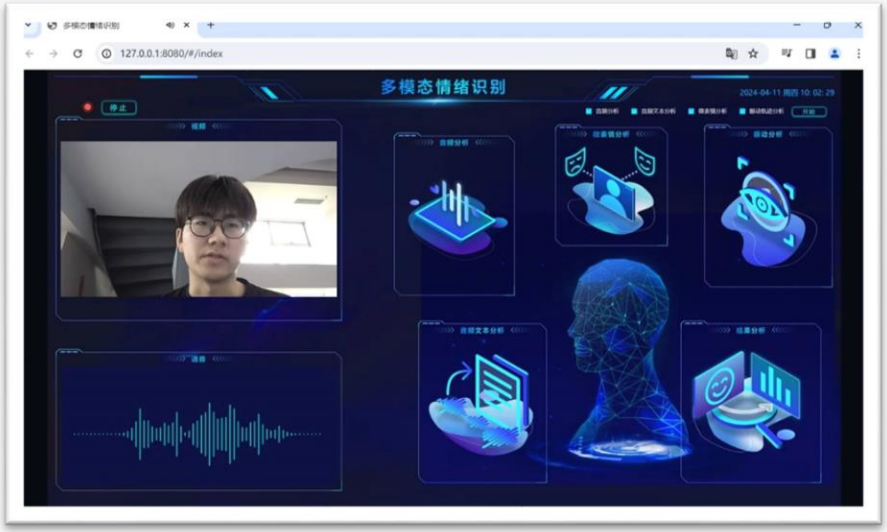
语音情感识别

多模态情感分析广泛使用的方式包括特征层融合、决策层融合与模型层融合，以文本、语音与图像三个模态为例。









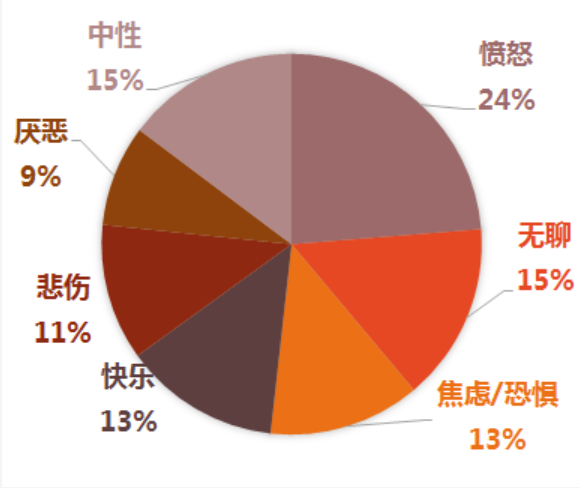
语音情感识别

- ✓ **ACCorpus：**清华大学和中科院心理研究所合作录制，包含5个子库，多模态、多通道的情感数据库、语音情感数据库、人脸表情视频数据库、人脸表情图像数据库等。其中语音情感数据库50位录音人，5类情感。

- ✓ **CASIA汉语情感语料库：**中科院自动化所录制，4位录音人（2男2女），有5类不同的情感，对500句文本进行演绎，最终保留9600句。

- ✓ **IEMOCAP数据集：**南加州大学SAIL实验室采集，多模态对话数据库，大约12h的试听数据，包括视频、语音、文本和面部动作捕捉信息，由2部分数据组成（按剧本演绎+即兴演绎），10个演员，9种情绪。

- ✓ **Emo-DB情感语音库：**德国柏林工业大学录制，包含535条语音。语料文本共有10条，涵盖7类情感，均为日常口语，无情感倾向，由表演者用不同的感情演绎出来。采样频率：48kHz（后压缩至16kHz）。



西安交通大学
XI'AN JIAOTONG UNIVERSITY

谢谢大家