中国石油大学(华东)
CHINA UNIVERSITY OF PETROLEUM

计算机科学与技术学院
College of Computer Science & Technology

第十二章 语音处理

人工智能课题组

智能科学与技术系

目录 CONTENTS

12.1语音的基本概念

- 语音是指人类通过发声器官发出的、具有一定意义的、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。
- 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~30K赫兹，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，在音频处理过程中可以忽略。

12.1语音的基本概念

• 语音四要素

- 声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。
- 音高指声波频率，即每秒钟振动次数的多少，频率高则音高，反之则低。
- 音强指声波振幅的大小。音的强弱是由发音时发音体振动幅度（简称振幅）的大小决定的，两者成正比关系，振幅越大则音越“强”，反之则越“弱”。
- 音长指声波震动持续时间的长短，也称为“时长”；发音体振动持续久，声音就长，反之则短。
- 音色指声音的特色和本质，也称作音质。

12.1语音的基本概念

• 语音处理

- 声音经过采样以后，在计算机中以波形文件的方式进行存储，这种波形文件反映了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和不同的说话人**。
- 语音识别的难点之一在于语音信号的复杂性和多变性。一段看似简单的语音信号包含**说话人**、**发音内容**、**信道特征**、**方言口音**等大量信息；
- 此外，这些信息相互组合在一起，又表达了**情绪变化**、**语法语义**、**暗示内涵**等更为丰富的信息。在如此众多的信息中，仅有少量的信息和语音识别相关，这些信息被淹没在大量信息中，因此充满了变化性。

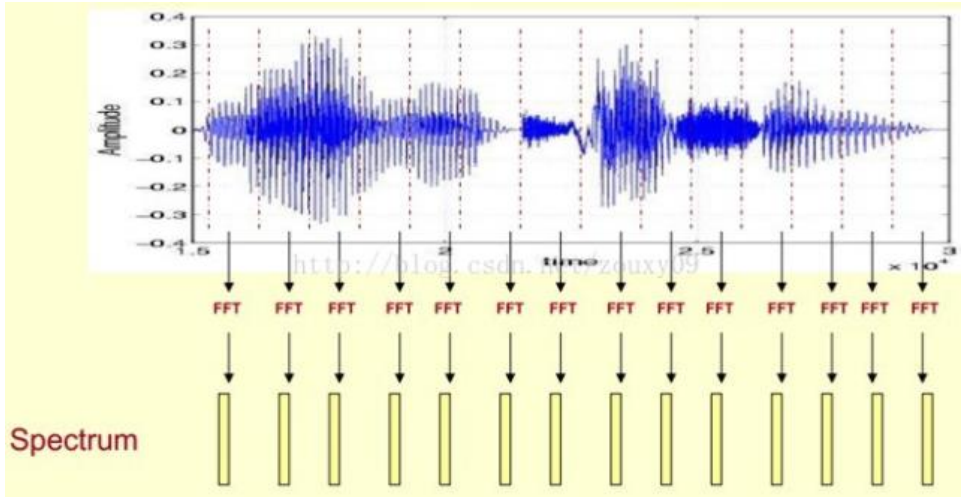
12.1语音的基本概念

• 语音识别的特征提取

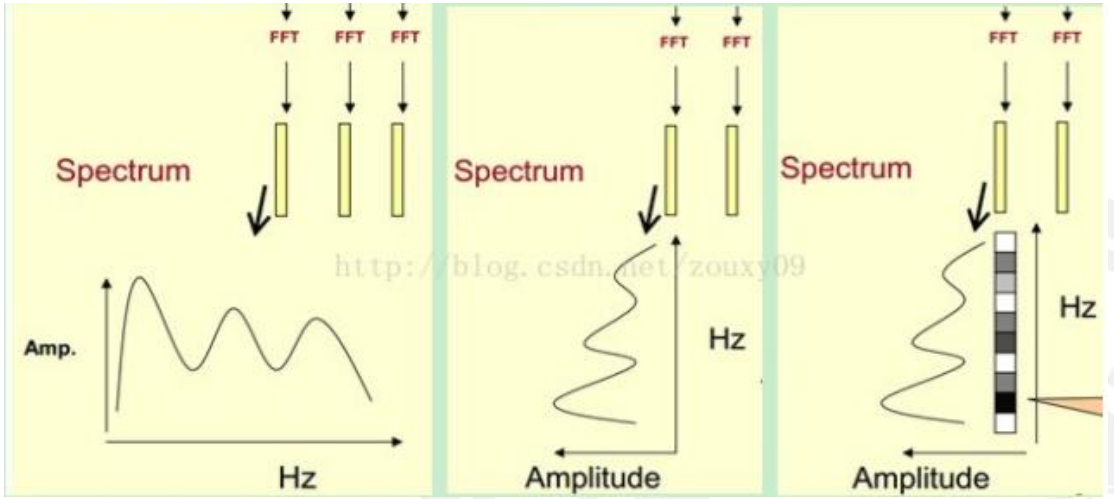
- 语音特征抽取即是在原始语音信号中提取出与语音识别最相关的信息，更好地反映不同语音的内容和音色差别，滤除其他无关信息。比较常用的声学特征有：**傅里叶谱**、**梅尔频率倒谱系数**、梅尔标度滤波器组特征和感知线性预测倒谱系数。
- 对语音进行频域上的转换，即提取语音频率的参数。例如，**傅里叶谱**、**梅尔频率倒谱系数**。
- 在此基础上，根据人耳的听觉特性，将语音信号在频域上划分成不同的子带，进而可以得到梅尔频率倒谱系数。梅尔频率倒谱系数是一种**能够近似反映人耳听觉特点**的频域参数，在语音识别和说话人识别上被广泛使用。

12.1语音的基本概念

- 语音识别的特征提取



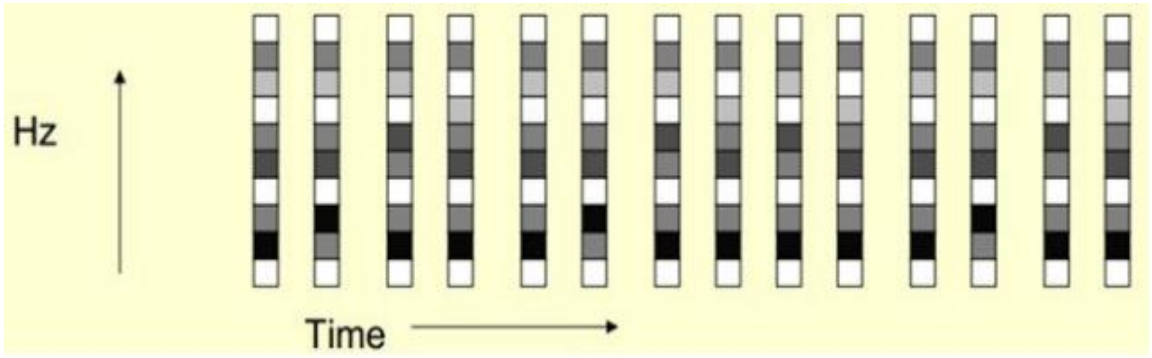
这段语音被分为很多短时帧（frame，大概10ms），每帧语音都对应于一个频谱（通过短时FFT计算），频谱表示频率与能量的关系。

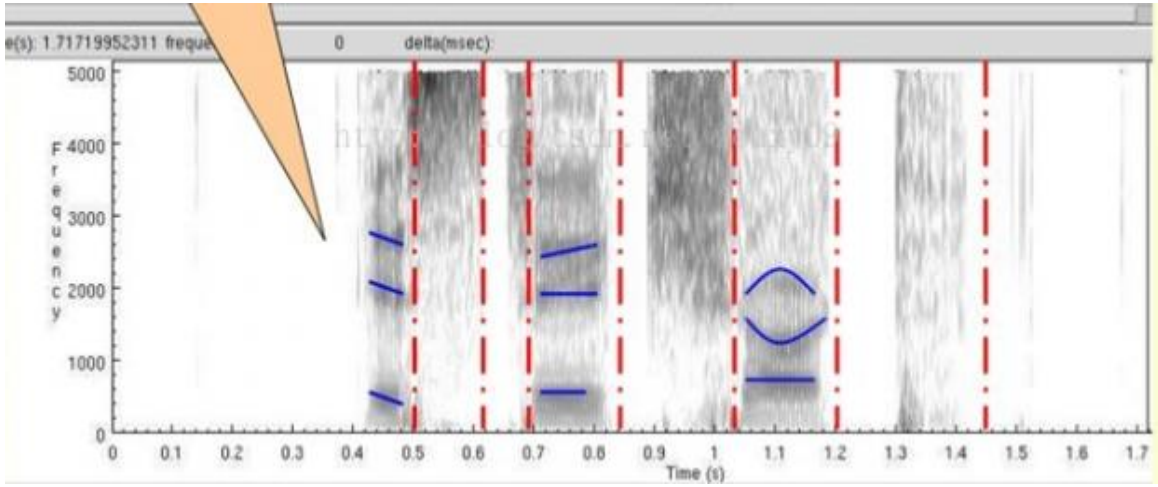


先将其中一帧语音的频谱通过坐标表示出来，如上图。现在我们将左边的频谱旋转90度。得到中间的图。然后把这些幅度映射到一个灰度级表示，0表示黑，255表示白色。幅度值越大，相应的区域越黑。这样就得到了最右边的图。

12.1语音的基本概念

- 语音识别的特征提取





这是要增加时间这个维度，这样就可以显示一段语音而不是一帧语音的频谱，这个就是描述语音信号的声谱图。可以直观的看到静态和动态的信息。

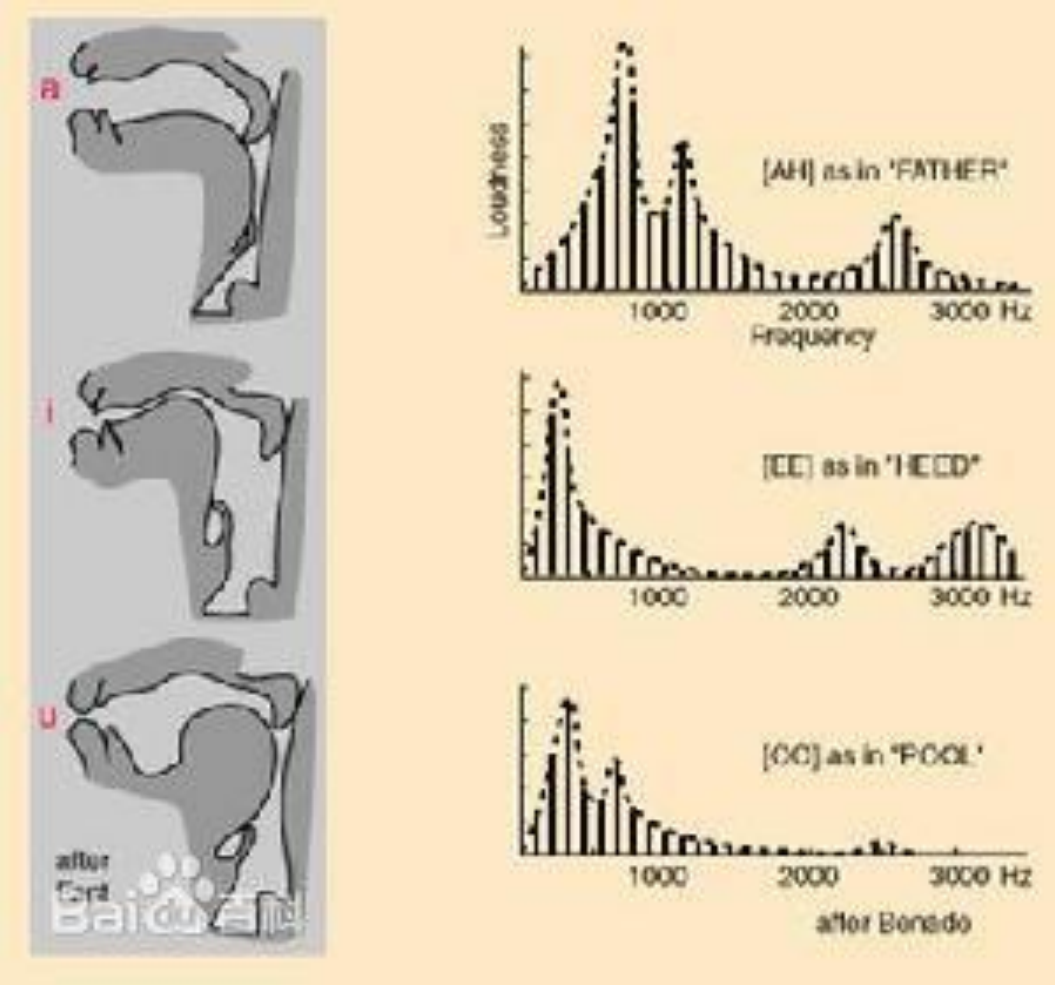
那我们为什么要在声谱图中表示语音呢？

首先，音素（Phones）的属性可以更好的在这里面观察出来。另外，通过观察共振峰和它们的转变可以更好的识别声音。隐马尔科夫模型就是隐含地对声谱图进行建模以达到好的识别性能。

还有一个作用就是它可以直观的评估TTS系统（text to speech）的好坏，直接对比合成的语音和自然的语音声谱图的匹配度即可。

12.1语音的基本概念

● 语音识别的特征提取



- 共振峰是指在声音的频谱中能量相对集中的一些区域，共振峰不但是音质的决定因素，而且反映了声道（共振腔）的物理特征。
- 音素（phone），是根据语音的自然属性划分出来的最小语音单位，依据音节里的发音动作来分析，一个动作构成一个音素。

12.1语音的基本概念

● 语音识别的特征提取

到现在，我们已有了频率曲线，但是人类听觉感知的实验表明，人类听觉的感知只聚焦在某些特定的区域，而不是整个频谱包络。

而Mel频率分析就是基于人类听觉感知实验的。

实验观测发现人耳就像一个滤波器组一样，它只关注某些特定的频率分量（人的听觉对频率是有选择性的）。也就说，它只让某些频率的信号通过，而压根就直接无视它不想感知的某些频率信号。但是这些滤波器在频率坐标轴上却不是统一分布的，在低频区域有很多的滤波器，他们分布比较密集，但在高频区域，滤波器的数目就变得比较少，分布很稀疏。

12.1语音的基本概念

- 语音识别的特征提取

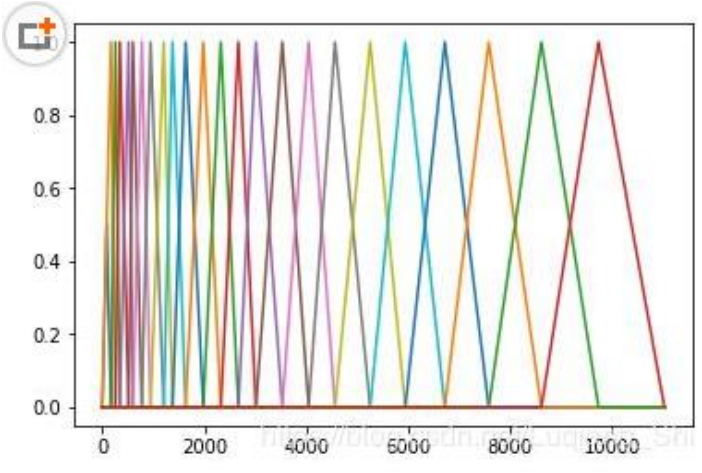


图8 Mel滤波器

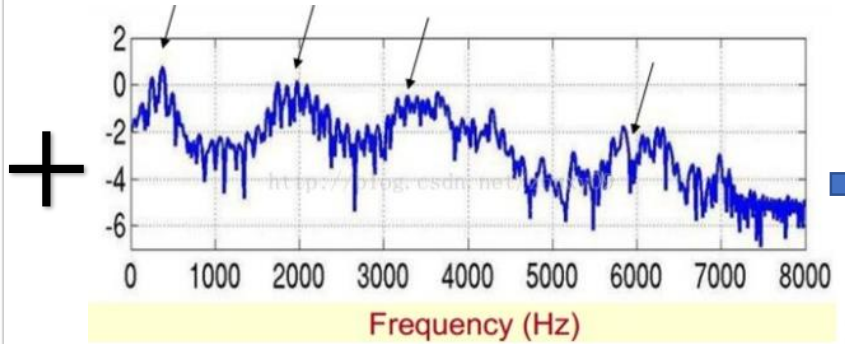


图9 原始语音频谱图

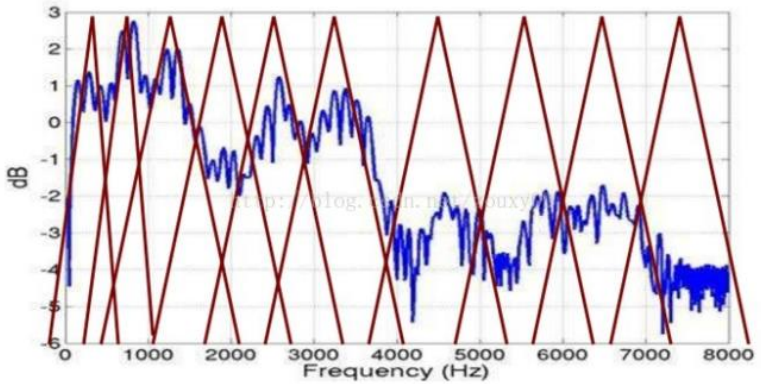
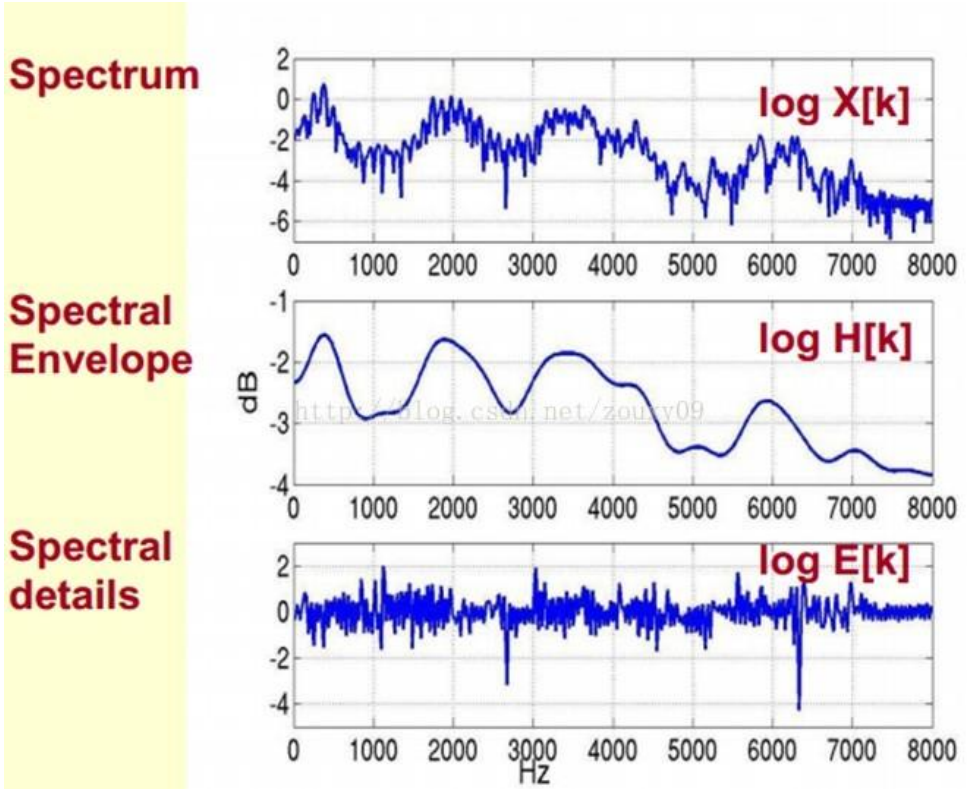


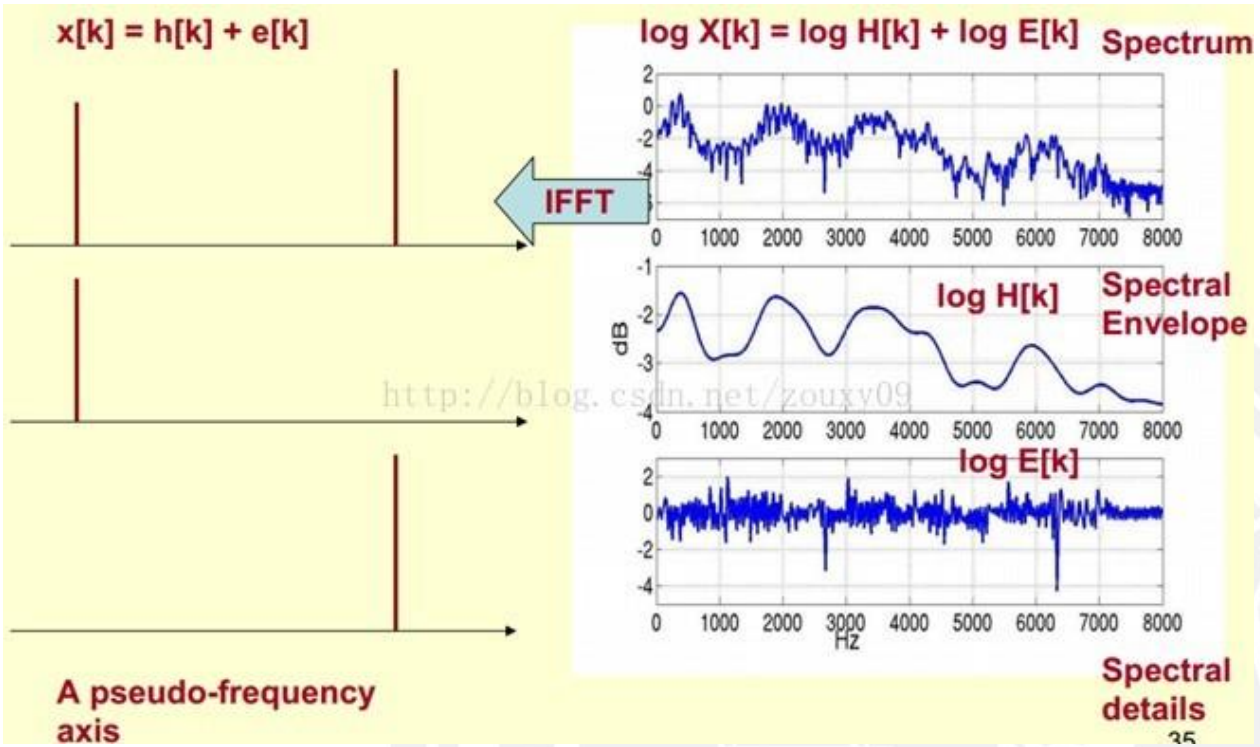
图10 梅尔频率图

在语音特征的提取上，人类听觉系统效果非常好，它不仅能提取出语义信息，而且能提取出说话人的个人特征，这些都是现有的语音识别系统所望尘莫及的。如果在语音识别系统中能模拟人类听觉感知处理特点，就有可能提高语音的识别率。

梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient，MFCC）考虑到了人类的听觉特征，先将线性频谱映射到基于听觉感知的Mel非线性频谱中。

12.1语音的基本概念-语音识别的特征提取





原始的频谱由两部分组成：包络和频谱的细节。包络是主要是低频成分（这时候的横轴可以看成时间）。而频谱的细节部分主要是高频。

在频谱上做傅里叶变换就相当于逆傅里叶变换Inverse FFT (IFFT)。在对数频谱上面做IFFT就相当于在一个伪频率（pseudo-frequency）坐标轴上面描述信号。

将x[k]通过一个低通滤波器就可以得到h[k]了。x[k]实际上就是倒谱Cepstrum。

12.1语音的基本概念

MFCC特征

- 总结下提取MFCC特征的过程：
  - 先对语音进行预加重、分帧和加窗；
  - 对每一个短时分析窗，通过FFT得到对应的频谱；
  - 将上面的频谱通过Mel滤波器组得到Mel频谱；
  - 在Mel频谱上面进行倒谱分析，获得Mel频率倒谱系数MFCC，这个MFCC就是这帧语音的特征；

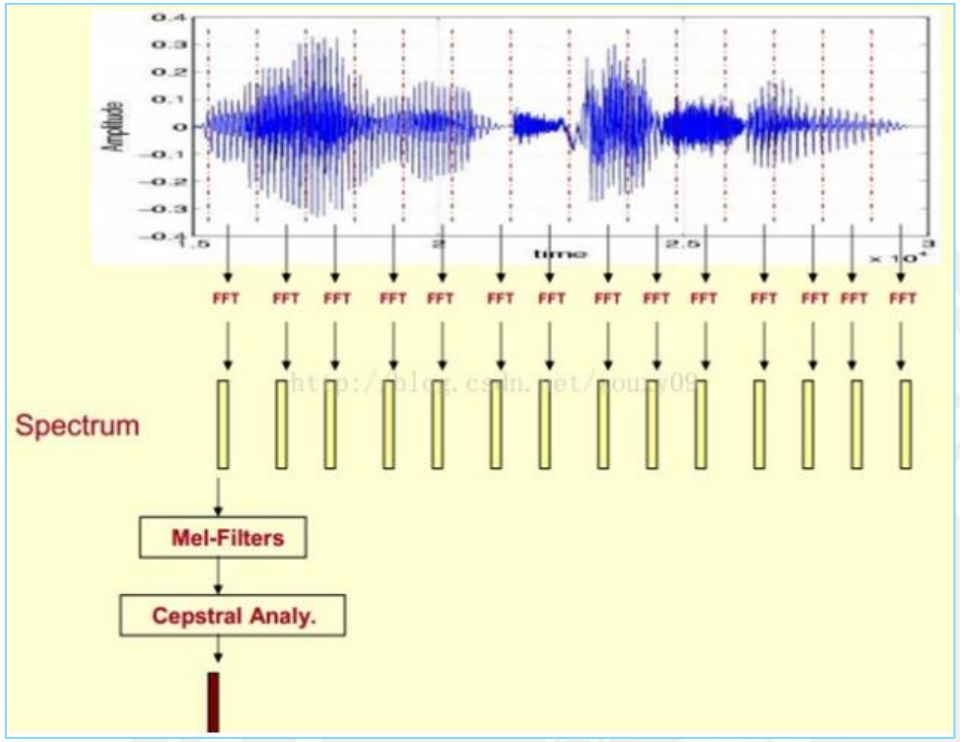


图11 MFCC特征提取过程

12.1语音的基本概念

● MFCC特征

- 至此，声音就成了一个12行（假设声学特征是12维）、N列的一个矩阵，称之为观察序列，这里N为总帧数。观察序列如下图所示，图中，每一帧都用一个12维的向量表示，色块的颜色深浅表示向量值的大小。

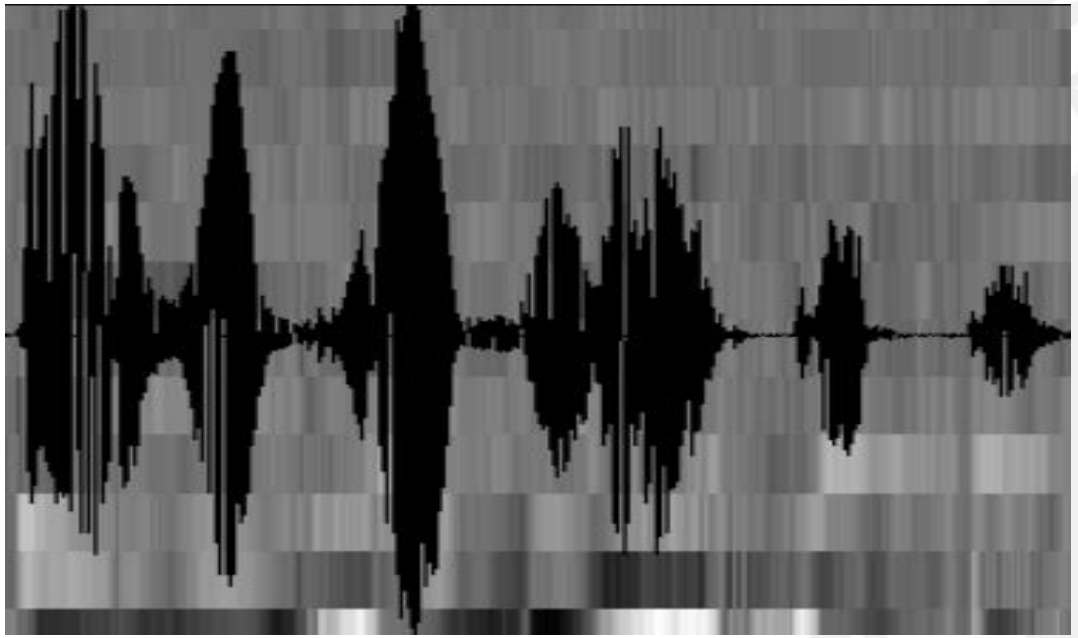


图12 MFCC声音特征图

12.1语音的基本概念

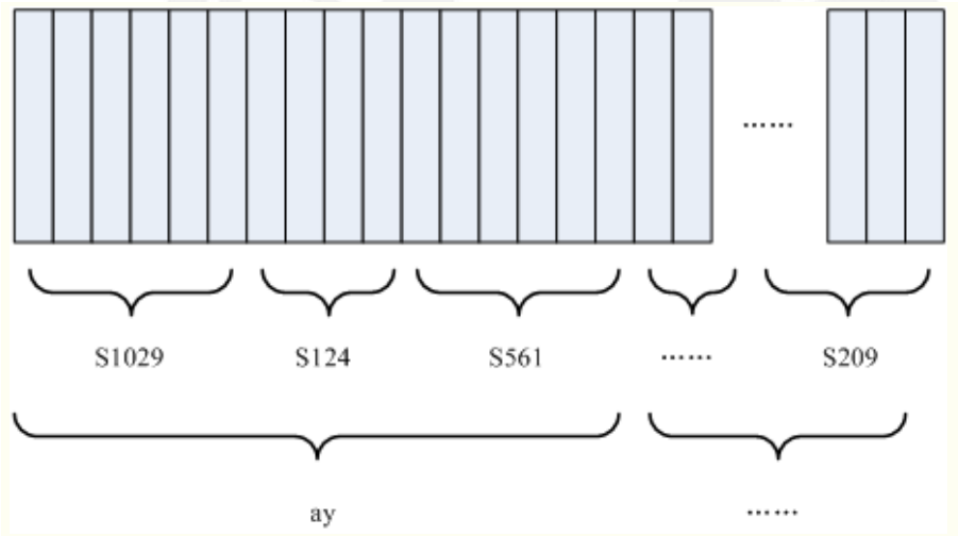
● 如何把MFCC特征矩阵变为文本

- 音素：单词的发音由音素构成。对英语，一种常用的音素集是卡内基梅隆大学的一套由39个音素构成的音素集，参见The CMU Pronouncing Dictionary。汉语一般直接用全部声母和韵母作为音素集。
- 状态：比音素更细致的语音单位。通常把一个音素划分成3个状态。
- 语音识别是怎么工作的呢？过程如下，见右图：

  第一步，把帧识别成状态（难点）；状态1:n帧

  第二步，把状态组合成音素；状态3:1音素

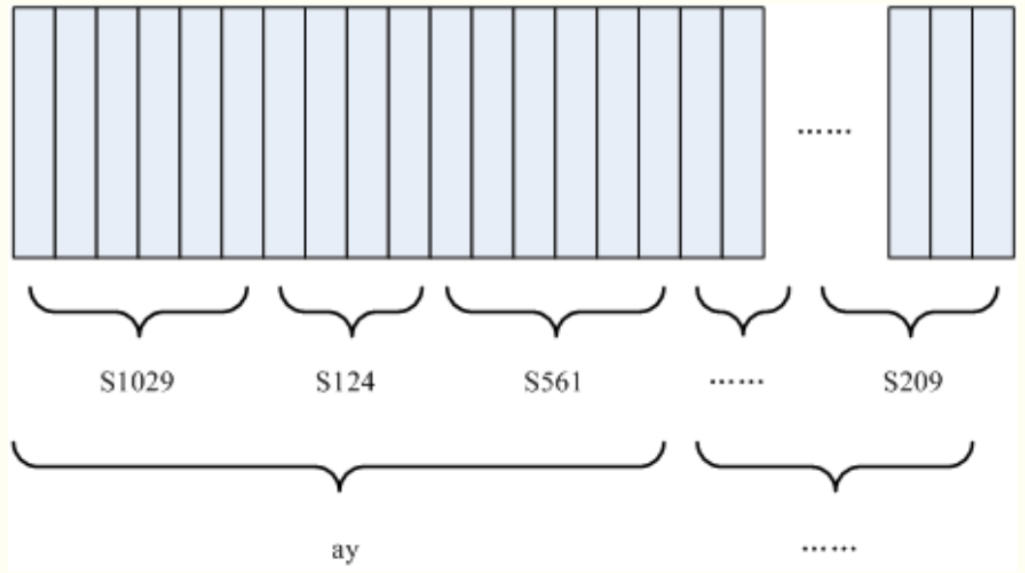
  第三步，把音素组合成单词。音素n:1单词



12.1语音的基本概念

● 如何把矩阵变为文本

- 每个小竖条代表一帧，
- 若干帧语音对应一个状态，
- 每三个状态组合成一个音素，
- 若干个音素组合成一个单词。
- 也就是说，只要知道每帧语音对应哪个状态了，语音识别的结果也就出来了。
- 那么第一个问题，如何将帧识别成状态呢？



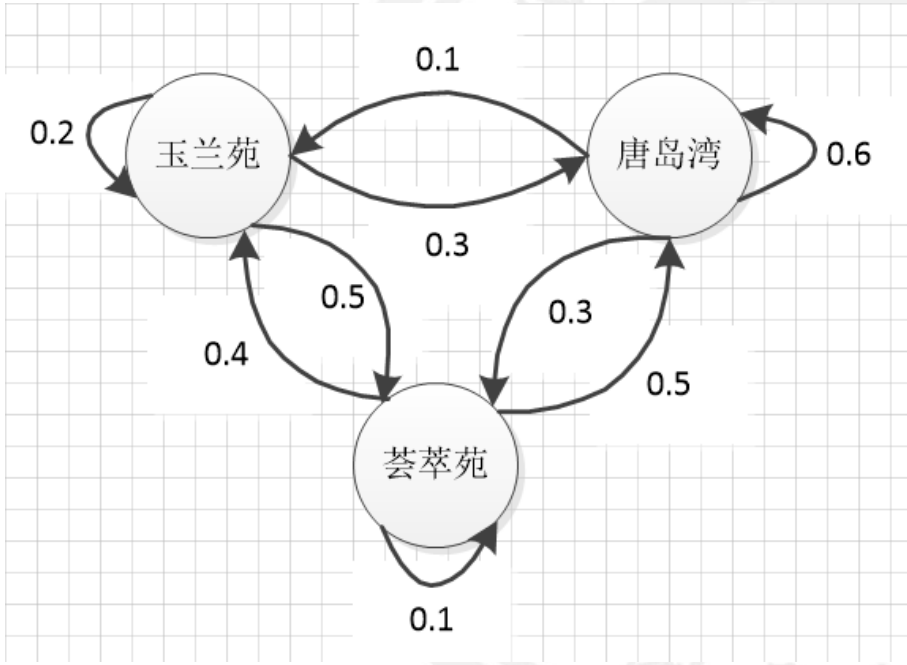
12.1语音的基本概念

- 马尔科夫链

大学生的我可以选择的吃饭地点就是这三个餐厅（真的太美味了！），n天以后我最可能在哪里吃饭呢？我自己都不知道，你知道吗？

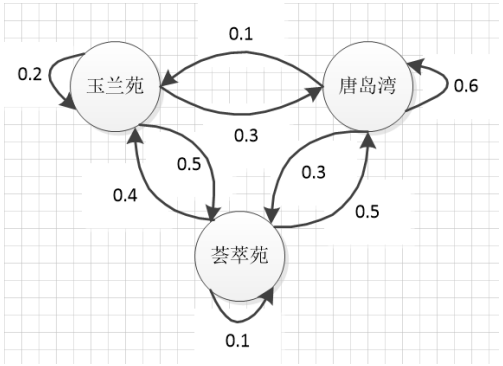
表1：状态转移矩阵

	玉兰苑	唐岛湾	荟萃苑
玉兰苑	0.2	0.3	0.5
唐岛湾	0.1	0.6	0.3
荟萃苑	0.4	0.5	0.1



12.1语音的基本概念

● 马尔科夫链



	玉兰苑	唐岛湾	荟萃苑
玉兰苑	0.2	0.3	0.5
唐岛湾	0.1	0.6	0.3
荟萃苑	0.4	0.5	0.1

- 这个矩阵就是转移概率矩阵，假设它是保持不变的，就是说第一天到第二天的转移概率矩阵跟第二天到第三天的转移概率矩阵是一样的。
- 有了这个矩阵，加上已知的第一天的状态分布，就可以计算出第N天的状态分布了。

每天中午12点　　在干嘛？

玉兰苑　唐岛湾　荟萃苑

1号　[0.6　0.2　0.2]　S1 → 已知

2号　[0.22]　S2 = S1*P

3号　S3 = S2*P

4号　S4 = S3*P

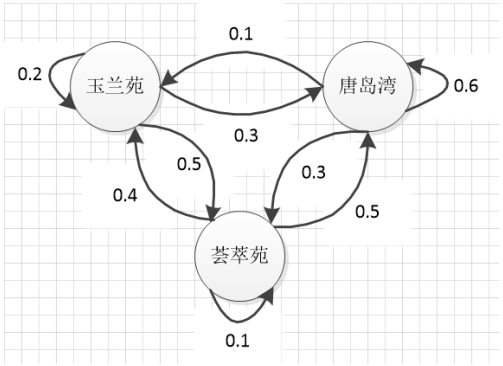
...

n号　[玉兰苑概率？唐岛湾概率？荟萃苑概率？]

状态分布矩阵S

12.1语音的基本概念

● 马尔科夫链



	玉兰苑	唐岛湾	荟萃苑
玉兰苑	0.2	0.3	0.5
唐岛湾	0.1	0.6	0.3
荟萃苑	0.4	0.5	0.1

- 这个矩阵就是转移概率矩阵，假设它是保持不变的，就是说第一天到第二天的转移概率矩阵跟第二天到第三天的转移概率矩阵是一样的。
- 有了这个矩阵，加上已知的第一天的状态分布，就可以计算出第N天的状态分布了。

每天中午12点　　在干嘛？

玉兰苑　唐岛湾　荟萃苑

1号　[0.6　0.2　0.2]　S1 → 已知

2号　[0.22　0.4　0.38]　S2 = S1*P

3号　S3 = S2*P

4号　S4 = S3*P

...

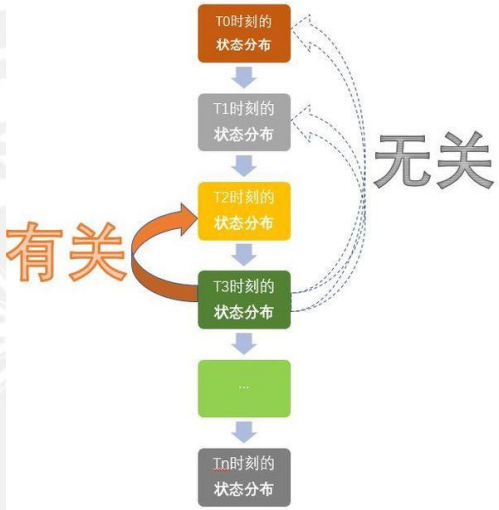
n号　[玉兰苑概率？唐岛湾概率？荟萃苑概率？]

状态分布矩阵S

12.1语音的基本概念

● 马尔科夫链

- S1是11月1号中午12点的状态分布矩阵［0.6，0.2，0.2］，里面的数字分别代表在玉兰苑的概率，唐岛湾的概率，荟萃苑的概率。那么：
- 11月2号的状态分布矩阵 S2 = S1 * P (:矩阵相乘)。
- 11月3号的状态分布矩阵 S3 = S2 * P (跟S1无关，只跟S2有关)。
- 11月4号的状态分布矩阵 S4 = S3 * P (跟S1，S2无关，只跟S3有关)。
- ...
- 11月n号的状态分布矩阵 Sn = Sn-1 * P (只跟它前面一个状态S_{n-1}有关)。
- 总结：
- 马尔可夫链就是这样一个的过程，它将来的状态分布只取决于现在，跟过去无关！知道了初始状态和转移矩阵就可以依次向后递推。



12.1语音的基本概念

● 隐马尔科夫链

- 有些时候人们不能直接观察到状态的值，即状态的值是隐含的，只能得到观测的值。为此对马尔科夫模型进行扩充，得到隐马尔科夫模型。

问:
张三去过哪家餐厅吃饭呀？好想也去体验一把。

12.1语音的基本概念

● 隐马尔科夫链

- 已有数据为：观测序列，但是还有状态转移矩阵以及观测矩阵，需要求出隐含状态转移序列。

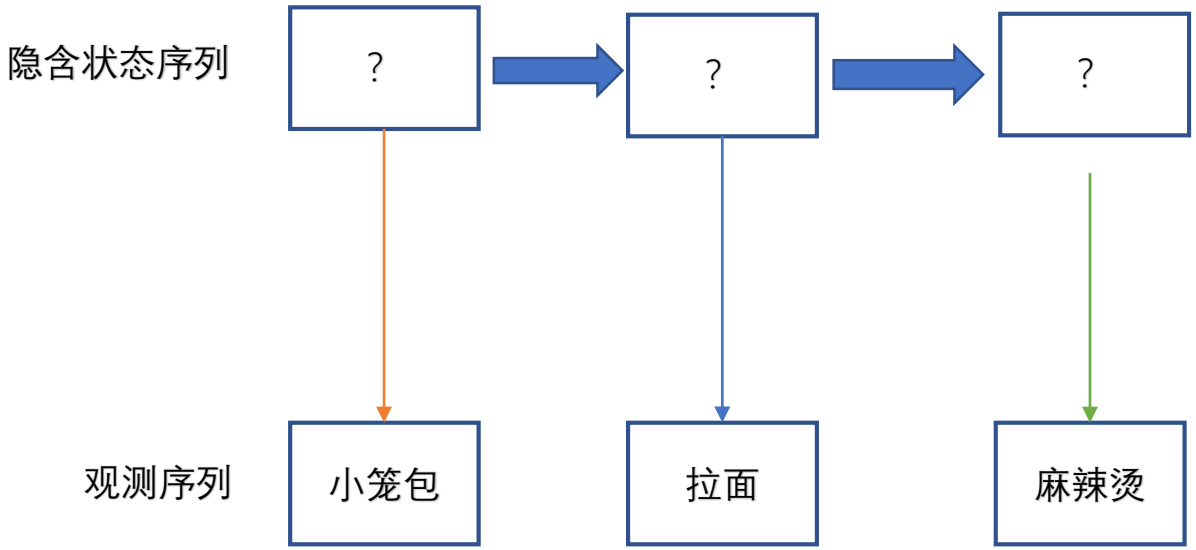


表1：隐含状态序列转移矩阵

	玉兰苑	唐岛湾	荟萃苑
玉兰苑	0.2	0.6	0.2
唐岛湾	0.3	0.4	0.3
荟萃苑	0.1	0.2	0.7

表2：观测矩阵

	小笼包	拉面	麻辣烫
玉兰苑	0.2	0.5	0.3
唐岛湾	0.1	0.2	0.7
荟萃苑	0.5	0.2	0.3

12.1语音的基本概念

● 隐马尔科夫链解决的关键问题

A：状态转移矩阵。B：观测矩阵。

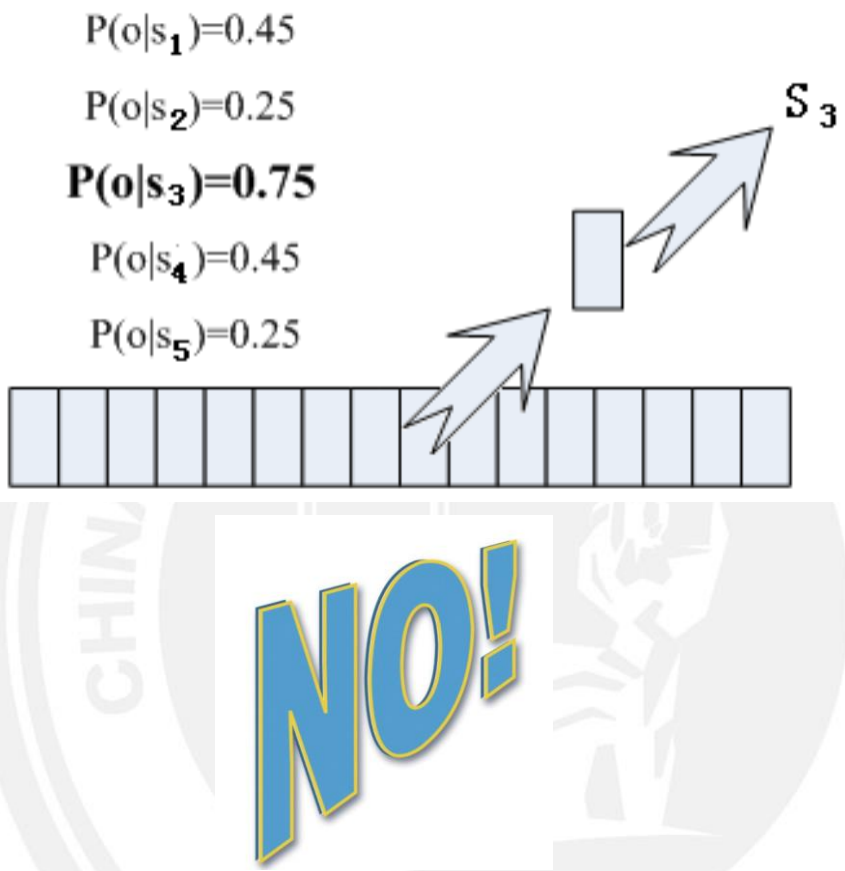
- 估值问题：给定隐马尔科夫模型的参数A和B，计算一个观测序列x出现的概率。
- 解码问题：给定隐马尔科夫模型的参数矩阵A和B以及一个观测序列x，计算最有可能产生此观测序列的状态序列。
- 学习问题：给定隐马尔科夫模型的结构，但参数未知，给定一组训练样本，确定隐马尔科夫模型的参数A和B。
- 语音识别就是其中的解码问题。

12.1语音的基本概念

● 帧与状态的转移图

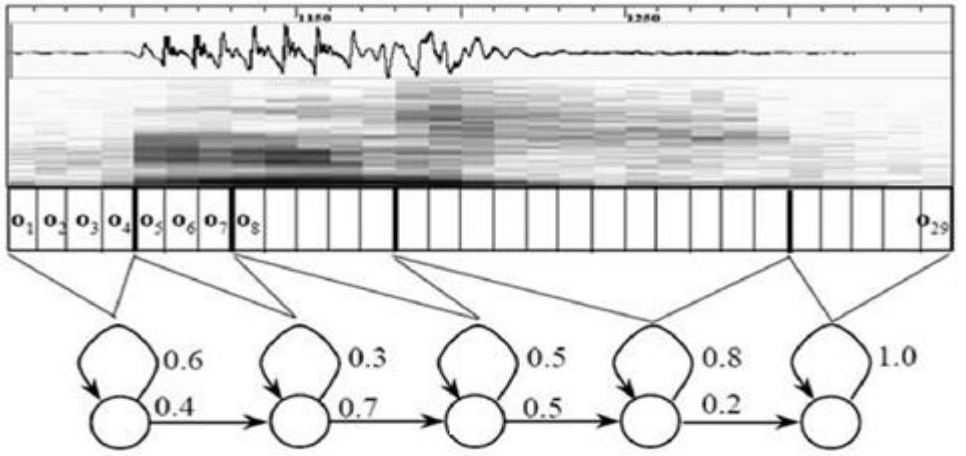
- 声学模型库中通过训练，存储了大量音素和帧之间的对应关系。
- 但是，每一帧都会得到一个状态号，最后整个语音就会得到一连串的状态号。
- 相邻帧的状态应该大多数都是相同的才合理，因为每帧很短。

- 如何将帧识别成状态呢？

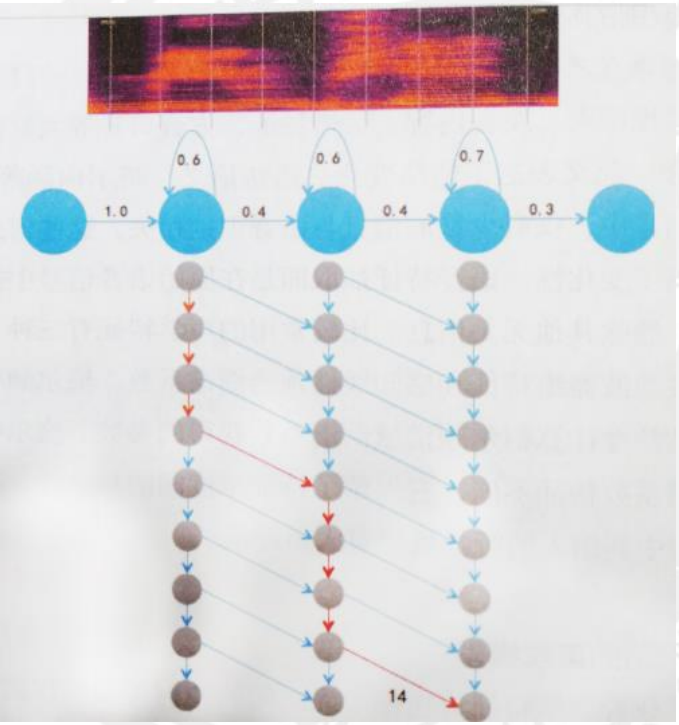


12.1语音的基本概念

● 帧与状态的转移图



- 解决这个问题的常用方法就是使用隐马尔可夫模型。
- 第一步，构建一个状态网络。
- 第二步，从状态网络中寻找与声音最匹配的路径。
- 这样就把结果限制在预先设定的网络中，比如你设定的网络里只包含了“我爱吃拉面”和“张三学习很好”两个句子的状态路径，那么不管说些什么，识别出的结果必然是这两个句子中的一句。
- 如果想识别任意文本呢？把这个网络搭得足够大，包含任意文本的路径就可以了。但这个网络越大，想要达到比较好的识别准确率就越难。所以要根据实际任务的需求，合理选择网络大小和结构。
- 搭建状态网络，是由单词级网络展开成音素网络，再展开成状态网络。语音识别过程其实就是在状态网络中搜索一条最佳路径，语音对应这条路径的概率最大，这称之为“解码”。



12.2 语音识别

- 语音识别是指将语音自动转换成文字的过程，在实际应用中，语音识别通常与自然语音、自然语言理解、自然语言形成及语音合成等技术相结合，提供一个基于语音的，自然流畅的人机交互系统。
- 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、解码搜索。语音识别系统的典型框架如下图所示：

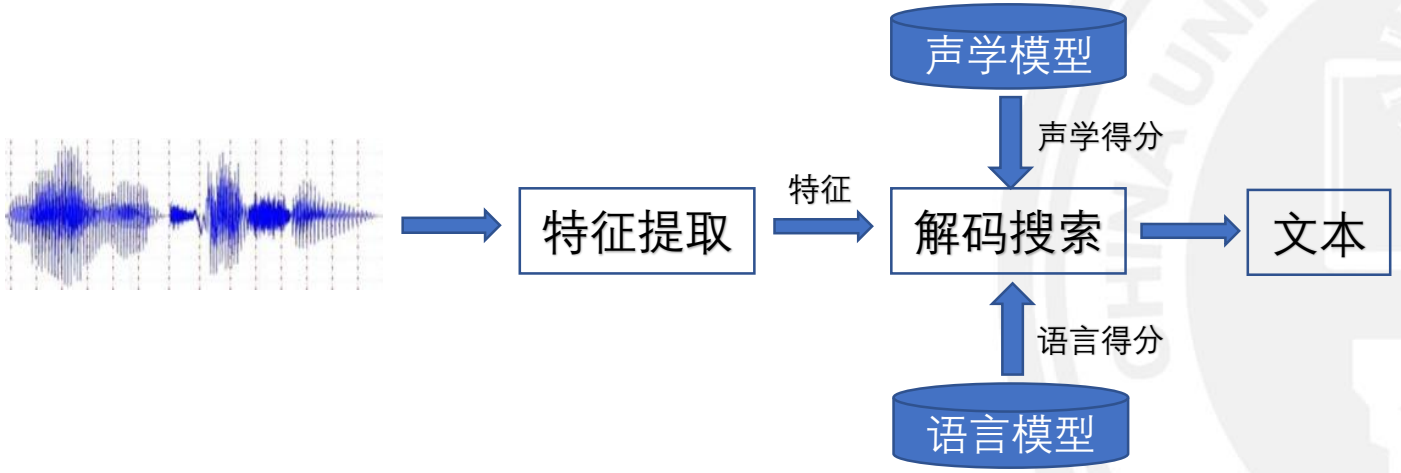


图14 语音识别系统框架

12.2.1 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**。一段看似简单的语音信号，其中包含了**说话人**、**发音内容**、**信道特征**、**方言口音**等大量信息；此外，这些信息互相组合在一起又表达了**情绪变化**、**语法语义**、**暗示内涵**等更为丰富的信息。在如此众多的信息中，仅有少量的信息与语音识别相关，这些信息被淹没在大量信息中，因此充满了变化性。
- 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**特征是指根据人耳听觉特征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**与梅尔频率倒谱系数特征不同，它保留了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。

12.2.2 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。在训练声学模型之前，需要选择建模单元，建模单元可以是**音素**、**音节**、**词语**等，其单元粒度依次增大。若采用词语作为建模单元，每个词语的长度不等，从而导致声学建模缺少灵活性；
- 此外，由于词语的粒度较大，很难充分训练基于词语的模型，因此一般不采用词语作为建模单元。相比之下，词语中包含的**音素是确定且有限的**，利用大量的训练数据可以充分训练基于音素的模型。因此，**目前大多数声学模型一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**，由于三因素的数量庞大，若训练数据有限，那么部分音素可能会存在信任不充分的问题，为了解决此问题，既往研究提出**采用决策树对三音素进行聚类以减少三音素的数目**。
- 比较经典的声学模型是混合声学模型，大致可以分为两种：**基于高斯混合模型-隐马尔可夫模型的模型**和**基于深度神经网络-隐马尔可夫模型的模型**。

12.2.3 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。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，通过语言模型可计算出其发生的概率P（今天天气怎么样）=0.8

  例2：令句子S=“材教智能人工”，该句子是病句，通过语言模型可计算出发生的概率P（材教智能人工）=0.000001

- 在语音识别系统中，语言模型所起的作用是在解码过程中从语言层面上限制搜索路径。常用的语言模型有**N元文法语言模型**和**循环神经网络语言模型**。
- 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。

12.2.4 语音识别的解码搜索

- 解码搜索的主要任务是在由声学模型、发音词典、语言模型构成的搜索空间内寻找最佳路径。需要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，一个词覆盖多帧语言特征，所以解码时声学得分和语言得分存在较大的数值差异。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尺度。

12.3 语音合成

- 语音合成也称为文语转换，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着广泛应用。主要步骤包括：

1. 文本预处理包含删除无效符号、断句等。
2. 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。
   - 它包含一系列步骤,依次是转换、清洗以及将文本数据标准化成可供 NLP、分析系统和应用程序使用的格式。
3. 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。

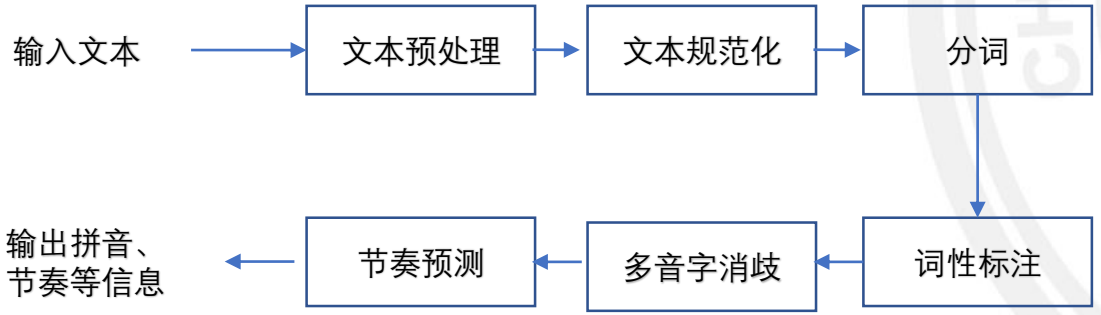


图15 汉语文本分析流程

12.3 语音合成

4. **字符转换**的任务是将待合成的**文字序列**转换为对应的**拼音序列**，即告诉后台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。
5. **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。直观的说，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。
   - 韵律表现是一个复杂现象，对韵律的研究涉及语言学、语音学、声学、心理学、物理学等多个领域。
   - 韵律模块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。与韵律相关的参数包括**基频**、**时长**、**停顿**和**能量**，韵律模型就是利用文本分析的结果来预测这四个参数。

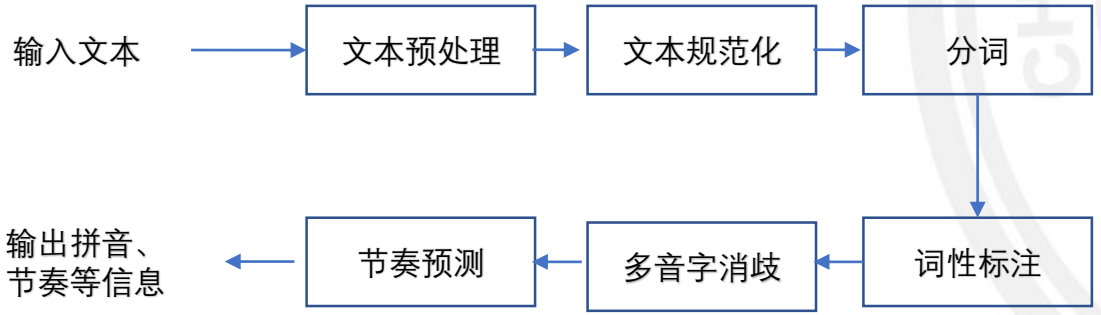


图15 汉语文本分析流程

12.3.1 基于拼接的语音合成方法

◆ 基本原理：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。

- 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。

✓ 随着计算机存储和运算能力的提升，实现基于大语料库的基元拼接合成系统成为可能。基元库由以前几MB扩大到几百MB甚至几GB，在音质和自然度上有了极大提高。

✓ 但是仍然存在：稳定性不够、拼接点不连续；难以改变发音特征、只能合成该建库说话人的语音。

12.3.2 基于参数的语音合成方法

- 主要分为**训练阶段**和**合成阶段**两个阶段，下图为系统框图。
- 在训练前，首先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**。
- 除了定义一些隐马尔科夫模型参数以及准备训练数据以外，模型训练前还有一个重要工作就是对**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语音参数有一定影响的上下文属性并设计相应问题。

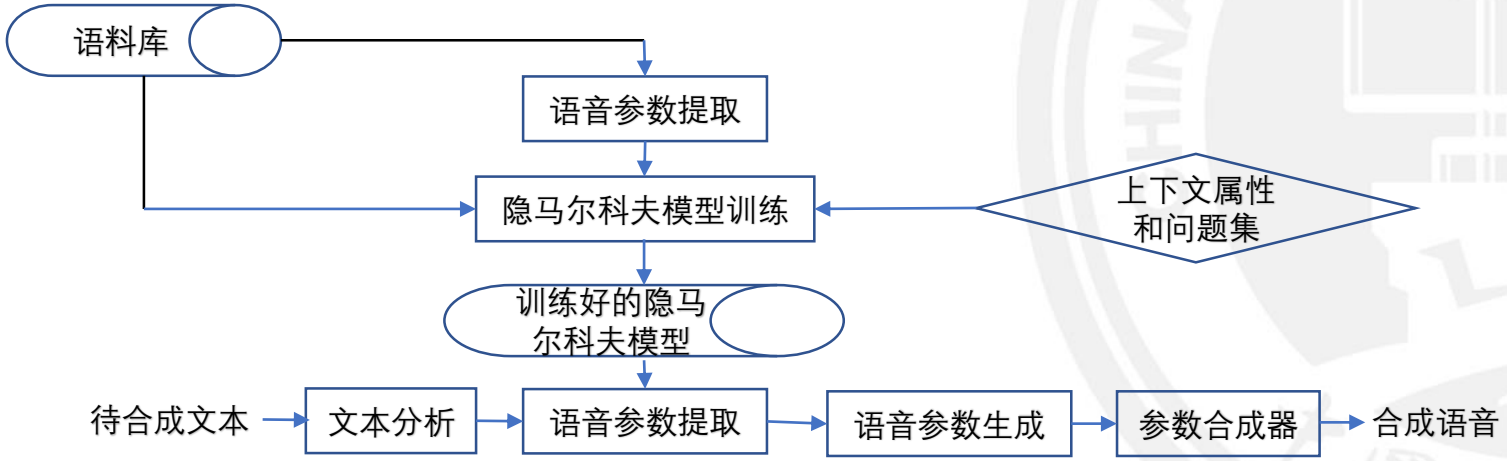


图16 基于隐马尔科夫模型语音合成系统框图

12.4 语音增强

- 语音增强一个最为重要的目标是实现释放双手的语音交互，通过语音增强有效抑制各种干扰信号、增强目标语音信号，使人机之间更自然交互。
  - ✓ 一方面可以提高话音质量，
  - ✓ 另一方面有助于提高声音识别准确性和抗干扰性。

◆ 定义：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。

12.4.1 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。

- 需要解决两个关键问题：

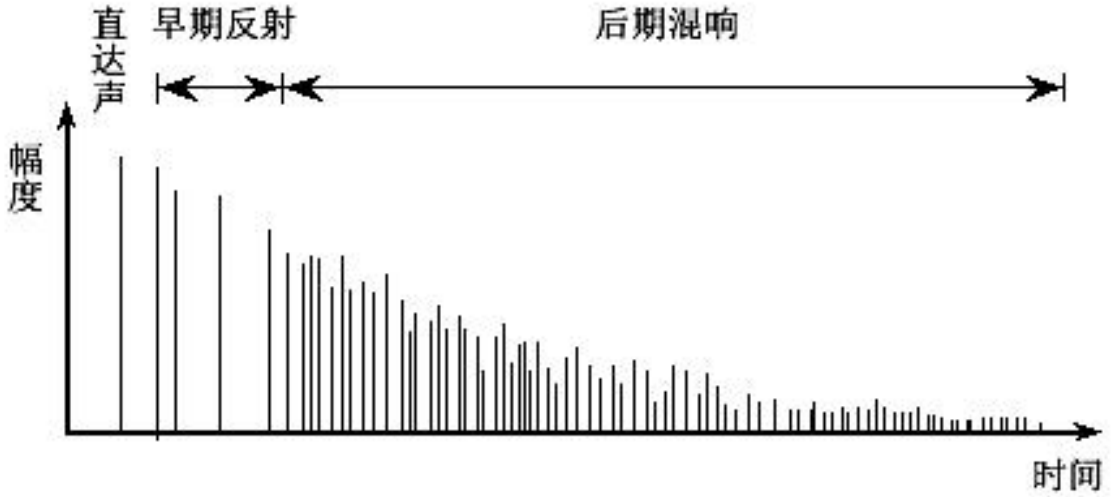
  1. 远端信号和近段信号的同步问题。

  2. 双讲模式下消除回波信号干扰的有效方法。

- 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，此时就需要有效的回声消除算法来抑制回声干扰，这在智能音箱、智能耳机中都是需要重点考虑的问题。

12.4.2 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声音的生成。



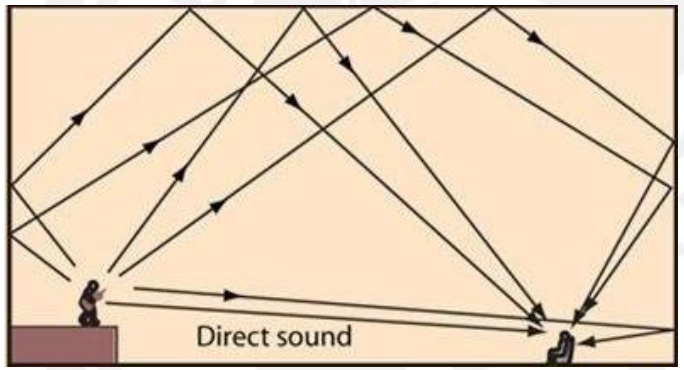


图17 KTV环境下混响干扰示意图

12.4.3 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。

➢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**。需要解决的核心问题是估计空间滤波器，它的输入是麦克风阵列采集的多通道语音信号，输出的是处理后的单路语音信号。由于声强和声音传播距离的平方成反比，因此，很难基于单个麦克风实现远场语音交互。**基于麦克风阵列的多通道语音降噪在远场语音交互中至关重要**。

➢多通道语音降噪算法通常受限于麦克风阵列的结构，随着麦克风个数的增多，噪声抑制能力会增强，但算法复杂度和硬件消耗也就相应增加，因此，**基于双麦的阵列结构**得到了广泛应用。

12.5 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人发出来的。
- 语音转换首先**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目标说话人的语音。实现一个完整的语音转换系统一般包括**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，然后根据某种匹配规则建立他们之间的匹配函数；
  - 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说话人的声音。

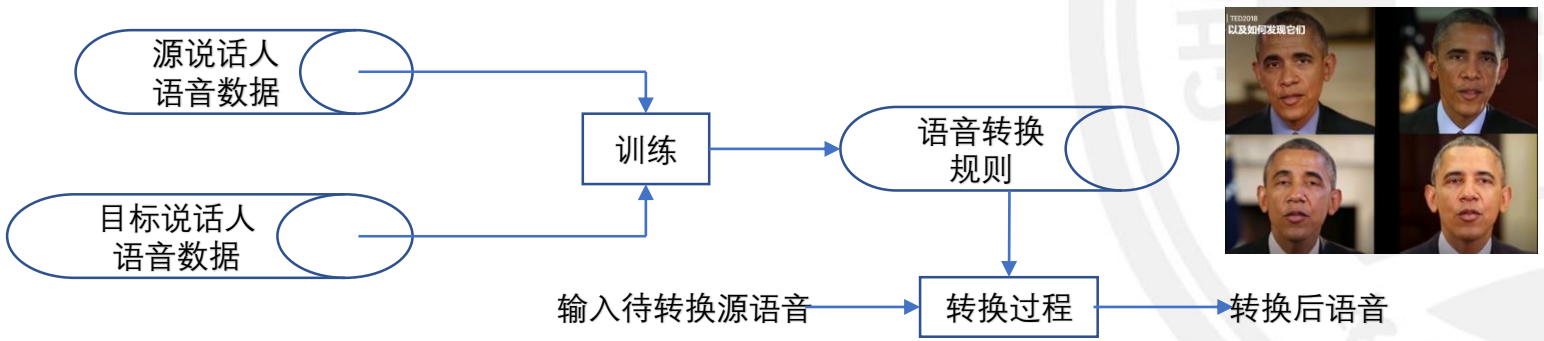


图18 语音转换基本系统框架

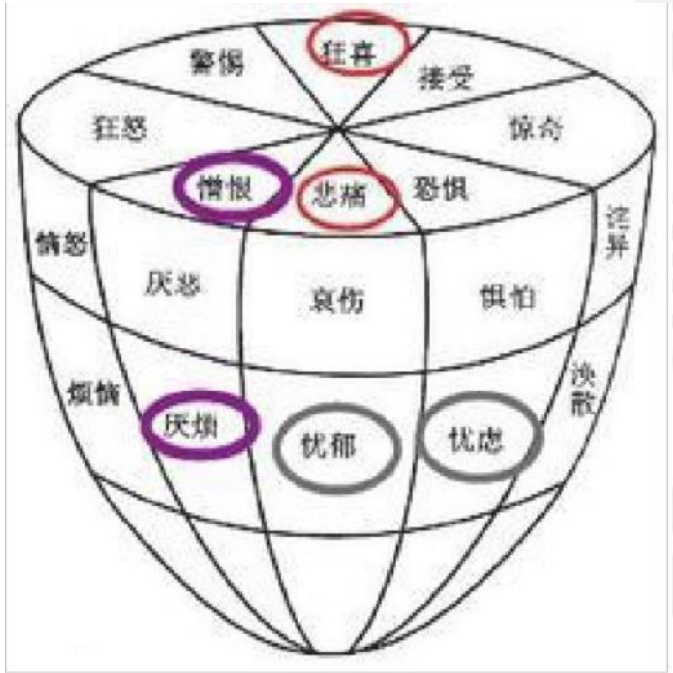
12.6 情感语音

- 研究语音信号的情感，首先要根据某些特征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。分析和处理语音信号中的情感信息、判断说话人的喜怒哀乐。
- 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，一般认为那些能够跨越不同人类文化，甚至能够为人类和具有社会性的哺乳动物所共有的情感类别为其基本情感，美国心理学家提出的六大基本情感：**生气、厌恶、恐惧、高兴、悲伤和惊讶。**
  - **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**属性，表明**情感正负面程度**的**愉悦度**属性等）。Russel等人在激活愉悦情感空间上，用一个情感轮对情感进行分类。

12.6 情感语音

• 情绪维度模型

- 三维的情感空间模型如下图所示，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。



12.6.2 情感语音的声学特征

- 情感语音中可以提取多种声学特征，用以反映说话人的情感行为特点。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**。
  - **韵律特征**具有较强的情感辨别能力，已经得到研究者的广泛认同。如**语速、能量和基频**等。
    - 在**激怒**状态时，语速比平常要快。
    - **喜、怒、惊**等情感的能量较大，而**悲伤**的情感的能量较低，这些能量差异越大，体现出的情感的变化也越大，
    - **欢快，愤怒和惊悸**语音信号的平均基频比较大，而悲伤的平均基频则较小。
  - 语音中所表达的情感状态被认为与**音质**有很大的相关性，语音特征主要有**呼吸声、明亮度特征和共振峰**等。
    - **欢快，愤怒，惊奇**和平静状态相比，振幅将变大；相反地，**悲伤**和平静相比，振幅将减小。
  - **频谱特征**主要包括线性谱特征和倒谱特征。

表12.1 情感和语音参数之间的关系

表1 情绪和语音参数之间的关系

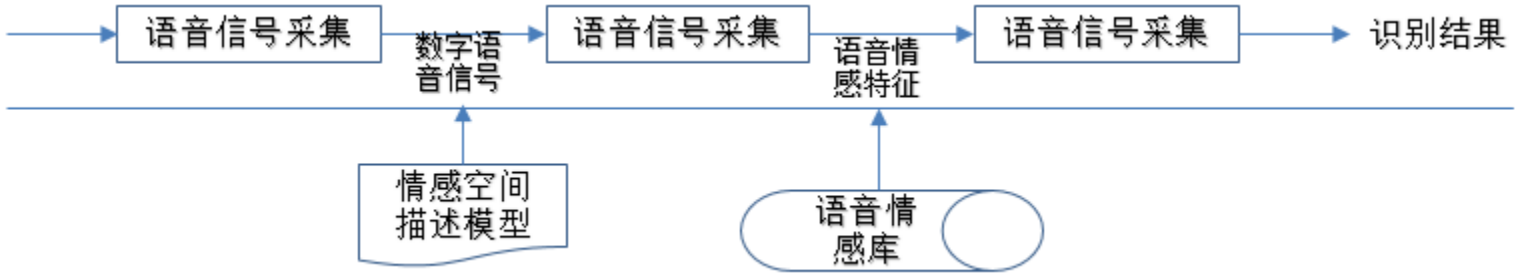
	愤怒	高兴	悲伤	恐惧	厌恶
语速	略快	快或慢	略慢	很快	非常快
平均基音	非常高	很高	略低	非常高	非常低
基音范围	很宽	很宽	略窄	很宽	略宽
强度	高	高	低	正常	低
声音质量	有呼吸声、胸腔声	有呼吸声、共鸣音调	有共鸣声	不规则声音	嘟囔声、胸腔声
基音变化	重音处突出	光滑、向上弯曲	向下弯曲	正常	宽、最终向下弯曲
清晰度	含糊	正常	含糊	精确	正常

12.6.3 语音情感识别

- 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。
- 一般来说，语音情感识别系统主要由三部分组成—**语音信号采集**、**语音情感特征提取**和**语音情感识别**。

1. **语音信号采集模块**通过语音传感器获得语音信号，并传递到语音特征提取模块；
2. **语音情感特征提取模块**对语音信号中的情感关联紧密的声学参数进行提取，
3. 最后送入**情感识别模块**完成情感判断。

- 需要指出的是，语音情感识别离不开**情感的描述**和**语音情感库**的建立。



THANKS