# 汉语语音

1. 汉语语音特点
   1. 音系简单，音素少，音节少
   2. 清辅音多，多是弱清音
   3. 有鲜明的轻重音和儿化韵
2. 汉语由音素构成声母和韵母，含有声调的韵母称为调母。汉语的一个音节就是一个字的音。汉语共有1332个有调音节。
3. 汉语有四种音节结构：V、CV、VC1、CVC1。C是除了[ng]外的全部辅音；V是单元音或复合元音；C1是鼻辅音[n]或[ng]。C称为声母，V或VC1称为韵母
4. 声母
   1. 无声——塞音和塞擦音特有
   2. 爆破——塞音和塞擦音特有
   3. 摩擦或噪声——清声母用噪声源；浊声母用噪音声源
   4. 送气——送气塞音和送气塞擦音特有
   5. 过渡——高度动态的浊音音段
5. 韵母
   1. 过渡
   2. 起始目标值
   3. 核心目标值——绝大多数音节的核心部位，具有典型的频谱模式
   4. 收尾目标值
   5. 鼻尾
6. 汉语韵母有三类：8个单韵母、14个复韵母、16个鼻韵母
7. 元音和辅音是按音素的发音特征分类的，声母韵母则是按音节结构分类的
8. 对于不送气塞音[b][d]，元音的音渡正是它们之间互相区别的主要音征
9. 汉语的声调有阴平、阳平、上声、去声以及“轻声”五种音调，承担着重要的构字辨义作用。
10. 声调变化就是浊音基音周期（基音频率）的变化。变化的轨迹称为声调曲线。声调曲线从一个韵母起始端开始，到韵母终止端结束。
11. 不同声调的声调曲线开始段称为弯头段，呈共同上升走向；末尾段呈共同下降走向，称为降尾段。中间一段称为调型段。弯头段和降尾段对声调的听辨不起作用
12. 从声门到嘴唇的呼气通道叫做声道，声道的形状主要由嘴唇、颚，基音频率的范围在60~450Hz左右。基音频率决定了声音频率的高低，频率大音调高

# 听觉模型

1. 在人耳中，由于外耳道的共振作用 ，会使声音得到10dB左右的放大。日常谈话中，鼓膜位移约10-8cm。外耳在对声音的感知中起到声源定位和声音放大的作用。
2. 中耳的作用是进行声阻抗的变换，三块听小骨在一定声强范围内对声音起到线性传递，在特强声时，听小骨进行非线性传递，对内耳起到保护作用。内耳主要构成是耳蜗，把声音通过机械变换产生神经发放信号。
3. 科蒂氏器官由耳蜗覆盖、外毛细胞和内毛细胞构成。毛细胞上部的微绒毛收到耳蜗内流体速度变化的影响，引起毛细胞膜两边的电位变化。
4. 人的听觉系统有两个重要特性，一是耳蜗对于声信号的时频分析特性；另一个是人耳的听觉掩蔽特性
5. 不同频率的声音产生不同的行波，其峰值出现在基底膜的不同位置上。频率较低时，基底膜震动幅度峰值出现在基底膜顶部附近；频率较高时，基底膜震动的幅度峰值出现在基底膜基部附近。
6. 如果信号是一个多频率信号，则产生的行波将沿着基底膜不同的位置产生最大幅度。基底膜不同部位的毛细胞有不同的电学特性与力学特性。因此具有不同的机械谐振特性和电谐振特性。
7. 听觉掩蔽现象是指，在一个强行好附近，弱信号变得不可闻，被掩蔽掉了，同时存在一个弱信号和一个强信号频率接近时，强信号会提高弱信号的听阈。
8. 对于同时掩蔽现象，掩蔽声愈强，掩蔽作用愈大；掩蔽声与被掩蔽声的频率靠的愈近，掩蔽效果愈显著。
9. 声音不同时出现也存在掩蔽作用，称为短时掩蔽，又分为后向掩蔽和前向掩蔽

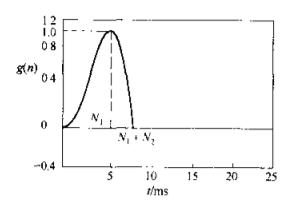
# 发音模型

通过对发音器官和语音产生机理的分析，可以将语音生成系统分为三个部分，在声门以下，负责产生激励振动，是“激励系统”；从声门到嘴唇的呼气通道是“声道系统”；嘴唇以外是辐射系统

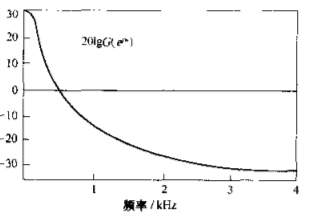
## 激励模型

激励模型分为浊音激励和清音激励，发浊音时由于声带不断张开和闭合，将产生间歇的脉冲波。这个脉冲波的波形类似于斜三角的脉冲





该三角波频谱为



由频谱可见三角波是低通滤波器

清音激励模拟成随机白噪声，一般使用均值为0的，方差为1的，在时间和/或幅值上为白色分布的序列。

## 声道模型

声道部分的数学模型有两种建模方法：

1. 把声道视为由多个等长的不同截面积的管子串联而成的系统，“声管模型”
2. 将声道视为一个谐振腔，“共振峰模型”

共振峰就是谐振频率

一般来说，元音使用前三个共振峰来表示，比较复杂的辅音或者鼻音，使用前五个以上共振峰。

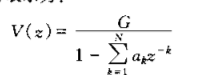
由成年人的声道长度可以推算共振峰频率，使用下面公式



基于物理声学的共振峰理论，有三个共振峰模型：级联型、并联型和混合型

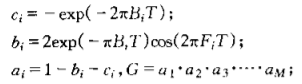
### 级联型

将声道看作一组串联的二阶谐振器。由于声道具有多个谐振频率和多个反谐振频率，所以可以被模拟为一个零点和极点的模型，零点表示反谐振频率，极点表示谐振频率。对于一般元音，谐振频率多于反谐震频率，因此表示为全极点模型



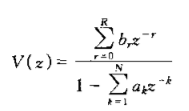
N为极点个数，G是振幅因子，ak是常系数。将其拆分为二阶极点的网络串联可以得到





### 并联型

对于非元音和大部分辅音，采用零极点模型。模型的传输函数为



通常来说，谐振频率多于反谐振频率，因此通常N>R。

假设，分子分母无公因子（没有谐振频率和反谐振频率相同），分母无重根。分解为并联模型可得到



### 混合型

级联的级数取决于声道长度，17cm对应取3~5级，为完整表示元音以及辅音，将级联型和并联型进行混合，根据要描述的语音自动切换

## 辐射模型

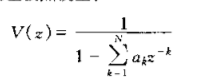
从声道模型输出的是速度波uL(n)，而语音信号是声压波pL(n)，二者之间的倒比为辐射阻抗ZL，表征口唇的辐射效应，若将其视为平板开槽辐射情况阻抗公式为





**辐射模型为高通滤波器，而激励模型中斜三角波模型为低通滤波器，因此在信号分析中，常用“预加重技术”。取样后插入一个高通滤波器，只剩下声道部分，便于声道参数分析，语音合成时进行“去加重”处理。**

**完整的语音信号模型为****，U(z)是激励信号，浊音的U(z)是声门脉冲(斜三角脉冲序列的z变换)，清音状态下U(z)是一个随机噪声的z变换**



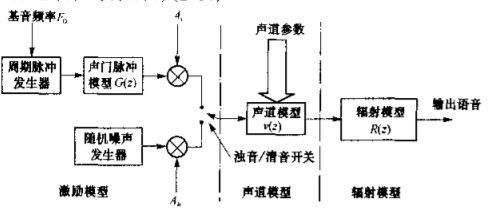
**R(z)可由zL(s)推出**





**R1较小，因此忽略极点得到**





**注意，这个模型是“短时”的模型，由于一些语音信号的变化缓慢，元音在10~20ms内的参数可假定不变**

# 语音信号统计特性

利用语音信号波形振幅概率密度函数和一些均值、自相关函数来表述统计特性。长时间范围内，大量取样数据的幅度绝对值计算幅度直方图，根据直方图寻找概率密度的表达式

振幅分布的概率密度逼近方法：

1. 修正伽马分布概率密度函数





1. 拉普拉斯分布概率密度函数





1. 高斯分布近似（正态分布）

# 语音信号分析

1. 虽然语音信号具有时变特性，但是在一个短时间范围内(10~30ms的短时间内)，其特性基本保持不变即相对稳定。任何语音信号的分析和处理必须建立在“短时”的基础上，将语音信号分为一段一段来分析其特征参数，每一段称为一帧，帧长一般取10~30ms。语音信号分析的结果是每一帧的特征参数组成的特征参数时间序列。
2. 根据分析的参数性质不同，可将语音信号分为时域分析、频域分析、倒频分析。根据分析方法不同，可分为模型分析和非模型分析方法。模型分析法是指根据语音信号产生的数学模型来分析和提取表征模型的特征参数，时域分析法、频域分析法以及倒频分析法都属于非模型分析法

## 语音信号数字化和预处理

数字化：放大及增益控制、返混叠滤波、采样、A/D变换及编码（PCM码）

预处理：预加重、加窗和分帧

在分析处理前要把分析的语音信号部分从输入信号中找出来，这项工作为语音信号端点检测。

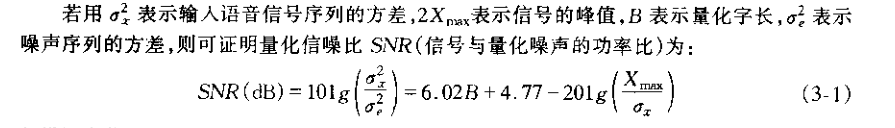
### 预滤波

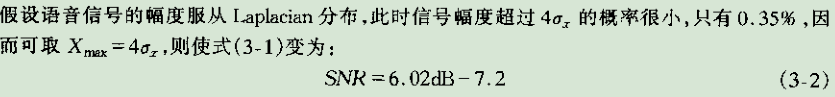
1. 抑制输入信号各频域分量中频率超出采样频率一半的所有分量
2. 抑制50Hz的电源工频干扰

一般来说，预滤波器的截止频率fH = 3400Hz，fL = 60~100Hz，采样率f = 8kHz。

### A/D变化

对信号进行量化，量化会产生误差，称为量化噪声。计算量化信噪比SNR（信号与量化噪声的功率比）的方法为：





注意这里的dB应该是印刷错误，应该是B，从上式可以得到每bit字长对SNR的贡献约为6dB。当B=7bit时，SNR=35dB。由于语音波形动态范围达55dB（上下约30dB），因此用附加5bit步长30dB左右的变化，因此用12bit来量化。

### 预处理

存储语音信号的方式一般为循环队列

由于声门激励和口鼻辐射的影响，高频端信噪比在800Hz以上不断跌落，因此要对高频部分进行预加重(Pre-emphasis)，使信号频谱变平坦，保持在低频到高频的频带中能用同样的信噪比求频谱。

预加重一般在语音信号数字化之后，使用具有6dB/倍频程的提升高频特征的预加重数字滤波器：



其中μ值接近于1，该滤波器的原理？

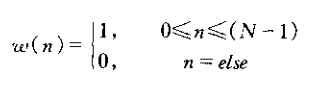
### 加窗分帧

一般每秒帧数为33~100帧。分帧时一般采用交叠分段的方法(为了使帧与帧之间平滑过渡)

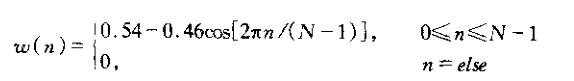
前一帧和后一帧交叠部分称为帧移，帧长和帧移的比值一般取0~1/2.分帧是由一定的窗函数w(n)来乘s(n)从而形成加窗语音信sw

窗函数：

1. 矩形窗



1. 汉明窗



窗函数的标准是：由于是在时域做变化，要减小时间窗两端的坡度，使窗口边缘两端不引起急剧变化而平滑过渡到0 ；在频域要有较宽的3dB（这个单位？）带宽，较小的边带最大值

### 窗的长度

频率分辨率和等于采用频率除以窗口长度即



该分辨率应该为时间分辨率（每个帧的时间长度）的倒数。当频率大于频率分辨率时，该频率将不可见。

对于时域分析来说，N太大使短时能量随时间变化小，不能真实地翻译语音信号的幅度变化；N太小，短时能量随时间有急剧变化，不能得到平滑函数。

窗口长度应考虑基音频率，通常在10kHz的采样频率下，N为100~200点。

## 语音信号的时域分析

### 短时能量及短时平均幅度分析

第n帧语音信号xn(m)的短时能量用En表示，计算公式为：



En是一个度量语音信号幅度值变化的函数，对高电平非常敏感，可采用另一个度量语音信号幅度值变化的函数，短时平均幅度函数Mn



En和Mn的作用是：

1. 可以区分浊音段与请音段，浊音的En比清音的En大
2. 区分声母和韵母的分界，有声和无声的分界，连字的分界
3. 作为超音段信息，用于语音识别

### 短时过零率分析

一帧语音中语音信号波形穿过横轴的次数。