



第二章 语音信号的数字模型

1

2.1 概述

2

2.2 语音的发音机理

3

2.3 语音的语音听觉机理

4

2.4 语音的感知

5

2.5 语音信号模型

6

2.6 语音信号数字模型



2.1 概述

本章重点介绍语音信号产生的数字模型，对语音信号的特性和听觉特性做一般介绍。



2.2 语音的发音机理

2.2.1 人的发音器官

1.组成

- (1) 肺和气管组成声源;
- (2) 喉和声带称为声门;
- (3) 由咽腔、口腔、鼻腔组成声道;

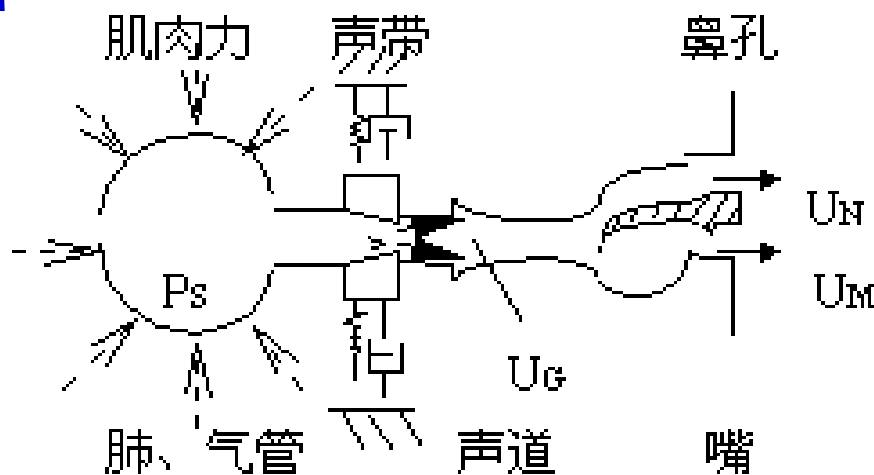


图2.1 发音器官机理模型



2. 功能

肺：产生压缩气体，通过气管传送到声音生成系统。

喉：控制声带运动的复杂系统。主要包括：环状软骨、甲状软骨、杓状软骨、声带。



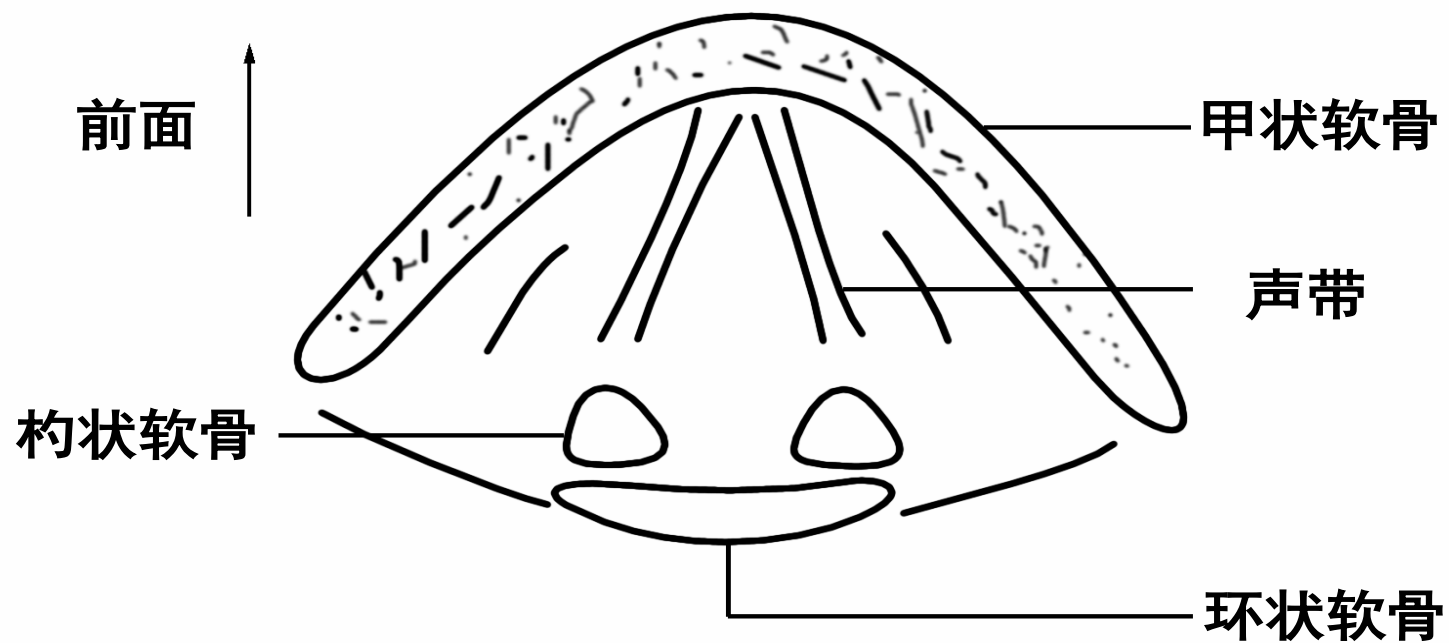


图 2.1 喉的平面解剖示意图





声门：声带之间的间隙称为声门。

主要功能：产生激励。

声道：声道指声门至嘴唇的所有发音器官。

包括：咽喉、口腔和鼻腔。

主要功能：传输调制声波。

声道的形状变化由舌、软腭、唇、牙决定。





口腔包括：上下唇、上下齿、上下齿龈、上下腭、舌和小舌等部分。

上腭又分为：硬腭和软腭两部分；

舌又分为：舌尖、舌面和舌根三部分。

鼻腔在口腔上面，靠软腭和小舌将其与口腔隔开。当小舌下垂时，鼻腔和口腔便耦合起来，当小舌上抬时，口腔与鼻腔是不相通的。**口腔和鼻腔都是发音时的共鸣器。**



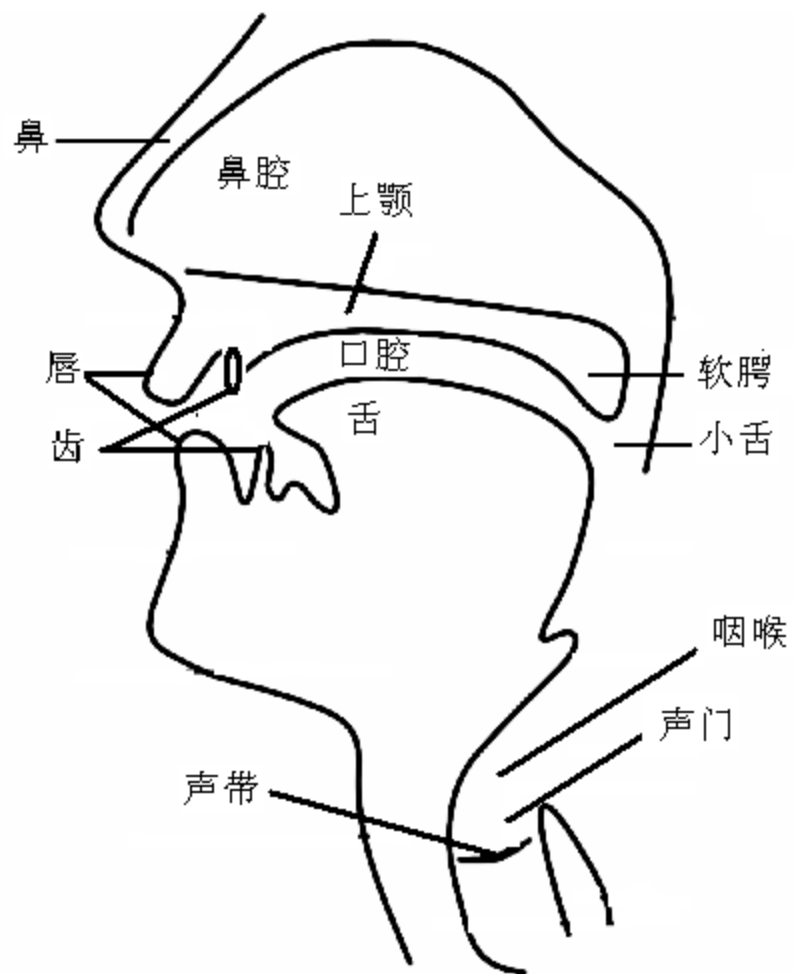


图2.3 声道纵剖面图





2.2.2 语音生成

图2.1为语音生成其机理模型。空气由肺部排入喉部，经过声带进入声道，最后由嘴辐射出声波，这就形成了语音。在声门（声带）以左，称为“声门子系统”，它负责产生激励振动；右边是“声道系统”和“辐射系统”。当发不同性质的语音时，激励和声道的情況是不同的，它们对应的模型也是不同的。



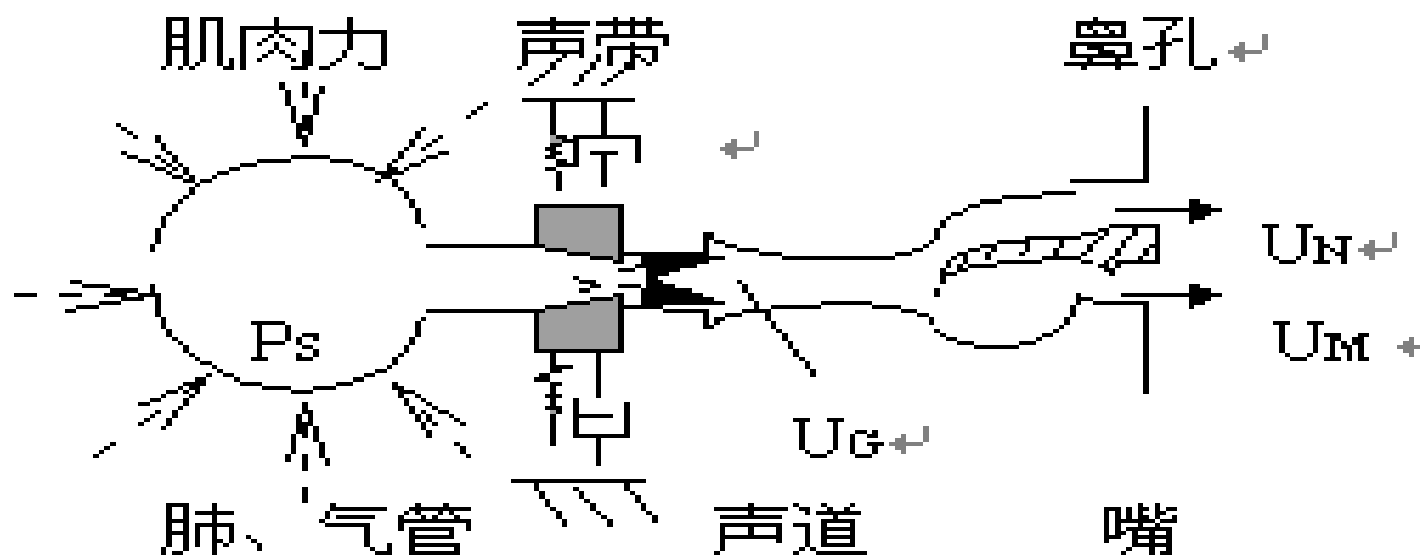
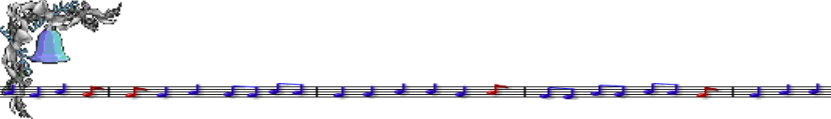


图 2.1 发音器官机理模型



语音生成动作可分为两种功能：

(1) 激励

(2) 调制

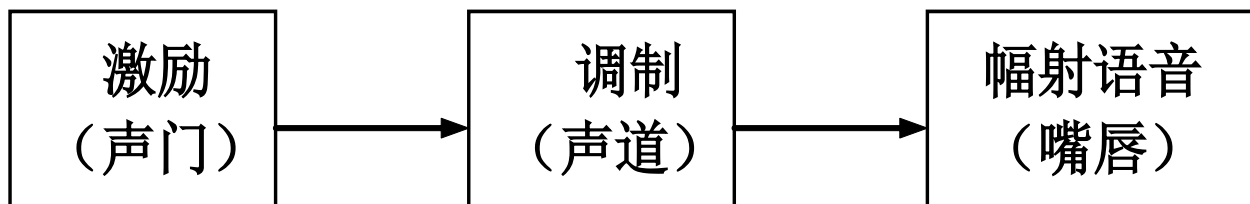


图 语音生成模型





2.2.2 语音生成-浊音

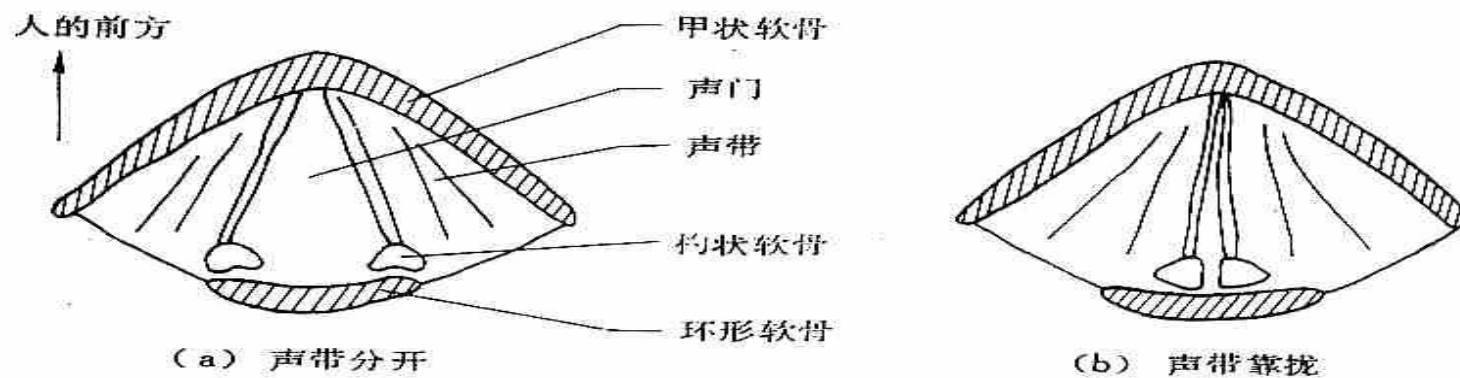
空气流经过声带时，如果声带是崩紧的，则声带将产生张弛振动，即声带将周期性地启开和闭合。声带启开时，空气流从声门喷射出来，形成一个**脉冲**，声带闭合时相应于脉冲序列的间隙期。因此，这种情况下在声门处产生出一个准周期脉冲状的空气流。该空气流经过声道后最终从嘴唇辐射出声波，这便是浊音语音。这个准周期脉冲的周期即为基音周期。





基音频率是由声带张开闭合的周期所决定的：
男性的基音频率一般为50~250Hz；
女性的基音频率一般为100~500Hz。





喉的解剖结构



典型的声门脉冲串波形





2.2.2 语音生成-清音

空气流经过声带时，如果声带是完全舒展开来的，则肺部发出的空气流将不受影响地通过声门。空气流通过声门后，会遇到两种不同情况。一种情况是，如果声道的某个部位发生收缩形成了一个狭窄的通道，当空气流到达此处时被迫以高速冲过收缩区，并在附近产生出空气湍流，这种湍流空气通过声道后便形成所谓**摩擦音**或**清音**。





2.2.2 语音生成-爆破音

另一种情况是，如果声道的某个部位完全闭合在一起，当空气流到达时便在此处建立起空气压力，闭合点突然开启便会让气压快速释放，经过声道后便形成所谓爆破音。





共振峰频率或共振峰

声音产生后，便沿着声道进行传播。声道可以看成是一根具有非均匀截面的声管，在发音时起着共鸣器的作用。声音进入声道后，其频谱必定会受到声道的共振特性的影响，声道具有一组共振频率，称为共振峰频率或共振峰。声道的频谱特性便主要地反映出这些共振峰的不同位置以及各个峰的频带宽度。共振峰及其带宽取决于声道的形状和尺寸，因而不同的语音对应于一组不同的共振峰参数。



2.3 语音的听觉机理

2.3.1 听觉器官

人的听觉器官包括：外耳、中耳和内耳

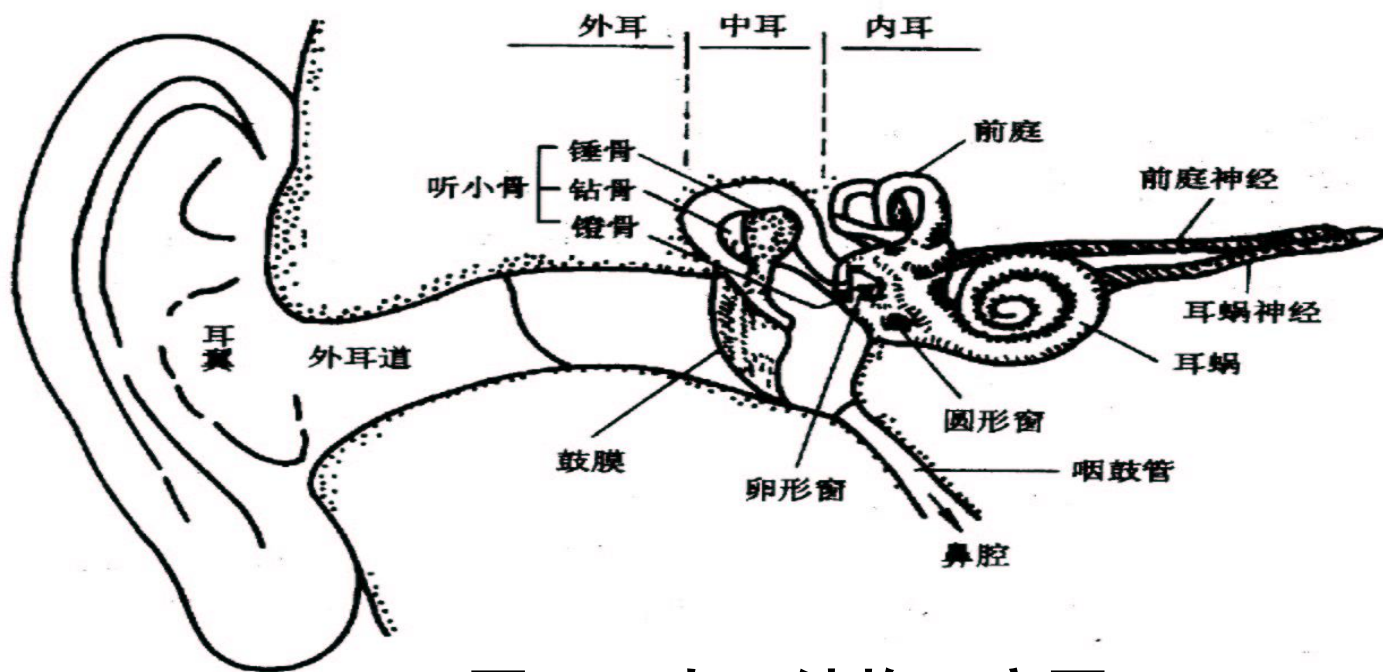


图2.3 人耳结构示意图



1. 外耳

外耳由耳廓(耳翼)、外耳道和耳鼓（鼓膜）组成。

2. 中耳

组成：包括三块听小骨:锤骨，砧骨和镫骨。

作用：阻抗匹配和限幅

外耳和中耳的综合作用相当于一个介于500Hz到6kHz之间的平滑的带通滤波器，可以用有限冲激响应(FIR---Finite Impulse Response)滤波器来模拟。





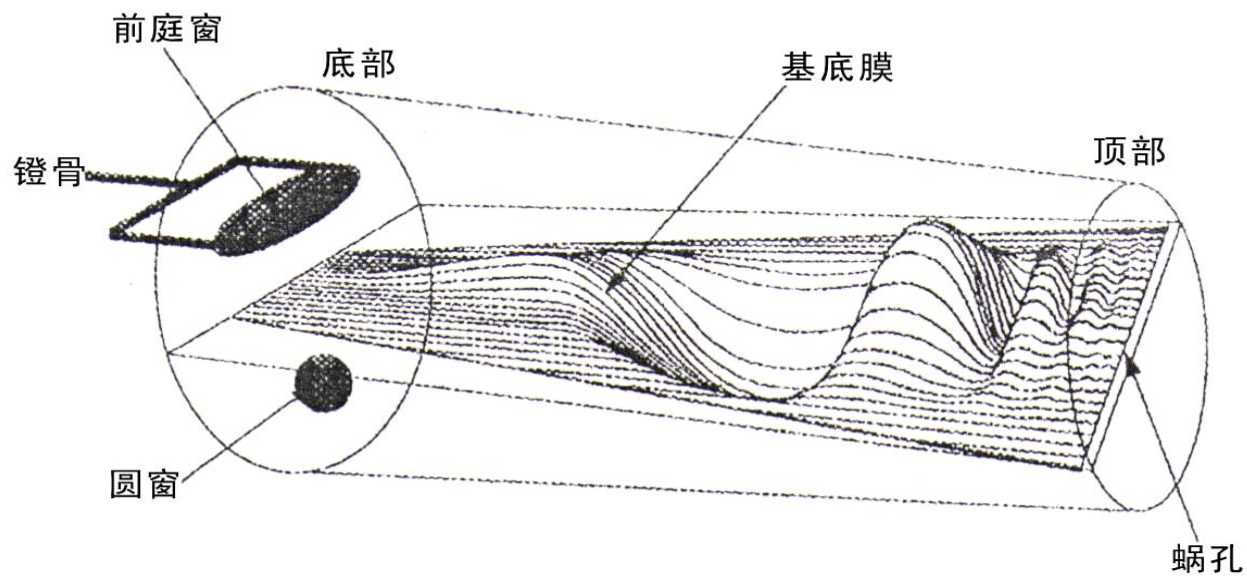
3. 内耳

内耳是一个充满液体的骨质结构，由前庭、圆形窗、卵形窗及耳蜗组成。





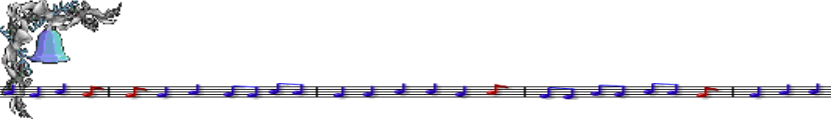
2.3.2 耳蜗的信号处理机制





当声音经外耳传入中耳时，镫骨的运动引起耳蜗内流体压强的变化，从而引起行波沿基底膜的传播。图2.6是流体波的简单表示。在耳蜗的底部基底膜的硬度很高，流体波传播的很快。随着波的传播，膜的硬度变得越来越小，波的传播也逐渐变缓。不同频率的声音产生不同的行波，而峰值出现在基底膜的不同位置上。





- 1 基底膜
- 2 内毛细胞
- 3 外毛细胞
- 4 听传导通路

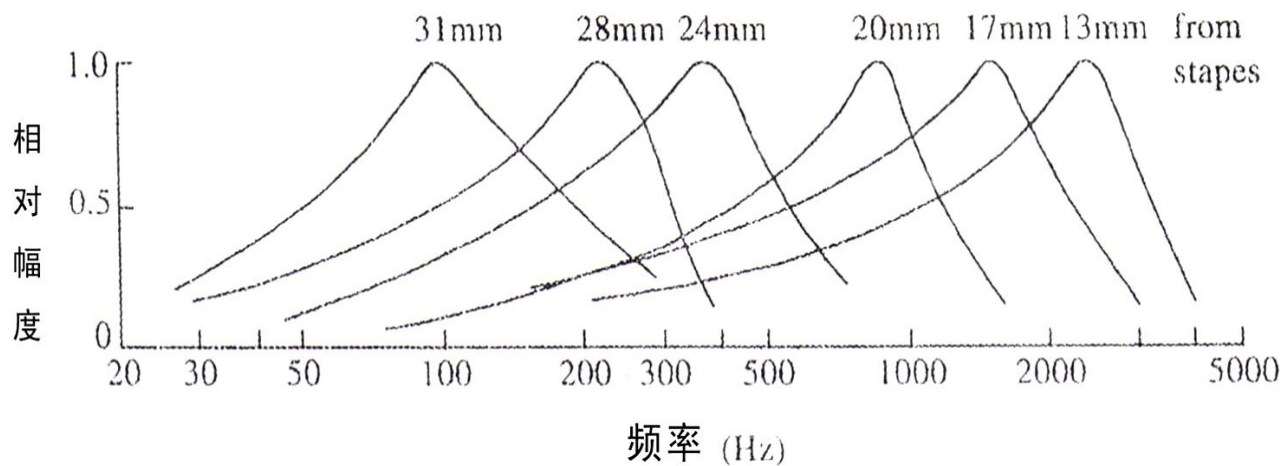


图2.7 基底膜上六个不同点的频率响应





2.3.3 语音信号听觉模型

听觉系统的研究主要集中在三个方面：听觉系统的实验研究、听觉系统的建模和听觉模型的应用。听觉系统的实验研究主要是指听觉系统在医学、生理学及心理学方面的研究。由于耳蜗深植于颅骨中，尺寸极小（如蜗管的直径只有1mm），所以耳蜗的实验研究是一项非常艰巨和复杂的工作。

耳蜗建模主要集中在基底膜的振动上，然而，建立基底膜的振动模型是耳蜗建模的首要任务，它又被称为耳蜗的宏观力学模型。



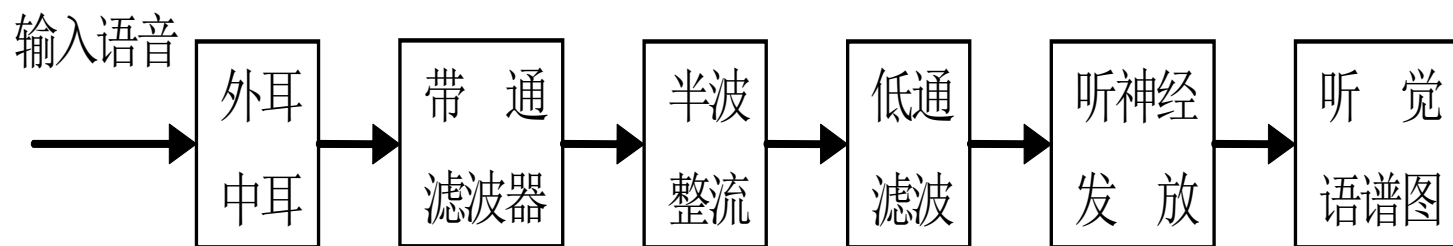
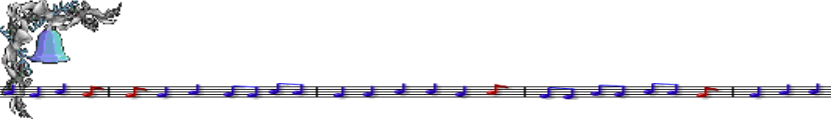


图2.10 语音信号听觉模型一般原理框图

语音信号首先通过一串带通滤波器(BPF)阵列，其中中心频率跟随着图2.7所示的基底膜频率响应按照对数尺度分布。每一个带通滤波器都被独立的设定为有限冲激响应滤波器(FIR)或无限冲激响应滤波器(IIR)，但是频率响应的波形并不是严格精确的。





被滤波的信号在通过内毛细胞/突触模型之后，到达听传导通路模型。虽然各种听觉模型的带通滤波器的性能特征是基本相同的，但是在接下来几级的信号处理过程却有很大差异。事实上，不同的听觉模型都各自拥有不同的IHC模型，突触模型和听传导通路模型。一些模型为每一个滤波后的信号都设有独立的频道，而另一些模型则认为在基底膜上相邻位置处滤波得到的信号之间存在耦合性。





根据人耳的听觉特性得出的模型作为语音识别的特征提取部分，可获得具有鲁棒性的特征参数，它们对真实世界中的噪音环境下的语音识别都表现出很好的性能。





2.4 语音的感知

2.4.1 几个概念

1. 人耳听觉界限的频率范围大约为20Hz-20kHz。
2. 语音感知的强度范围是0—130dB声压级。
3. **响度** 这是频率和强度级的函数。通常用响度(单位为宋)和响度级(单位为方)来表示。此时响度级定为零方。测量表明听阈值是随频率变化的。通常，人们把1kHz纯音听阈值定为零方。





4. 人耳刚刚可以听到的声音强度，称为“**听阈**”。

加大声音的强度，使听起来令耳朵感到疼痛，这个阈值称为“**痛阈**”。

5. **音高(音调)** 音高也叫基音。

物理单位为Hz，主观感觉的音高单位是Mel。当声强级为40dB频率为1kHz时，设定的音高为1000Mel。





2.4.2 掩蔽效应

掩蔽效应：

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，即：一个声音的听觉感受性受同时存在的另外一个声音的影响，这个现象称为人耳的“掩蔽效应”。此时前者称为被掩蔽音，后者称为掩蔽音。在掩蔽情况下，被隐蔽音的听阈会提高，即加大被掩蔽音的强度才能听到。此时听阈称为掩蔽听阈。



低频的纯音可以有效地掩蔽高频的纯音。

利用人耳的掩蔽效应，在进行语音压缩时，让量化噪音的频谱跟随语言信号频谱包络变化。则共振峰的频率成分就会掩蔽掉量化噪声。这个技术称为噪声整形或听觉加权处理。

低音容易掩蔽高音，而高音掩蔽低音较难。

基于此，可以将真实的声音频率映射到“感知”频率尺度，即Bark尺度对应的临界带宽。





2.5 语音信号模型

有三部分作用施加在语音的声波上：

声门产生的激励模型 $G(z)$ ；

声道产生的调制函数 $V(z)$ ；

嘴唇产生的辐射函数 $R(z)$ 。

语音信号的传递函数由这三个函数级联而成，

即： $H(z)=G(z)V(z)R(z)$



2.5.1 激励模型

发浊音时，产生的脉冲类似于斜三角形的脉冲。
激励波是一个以基音周期为周期的斜三角脉冲串。

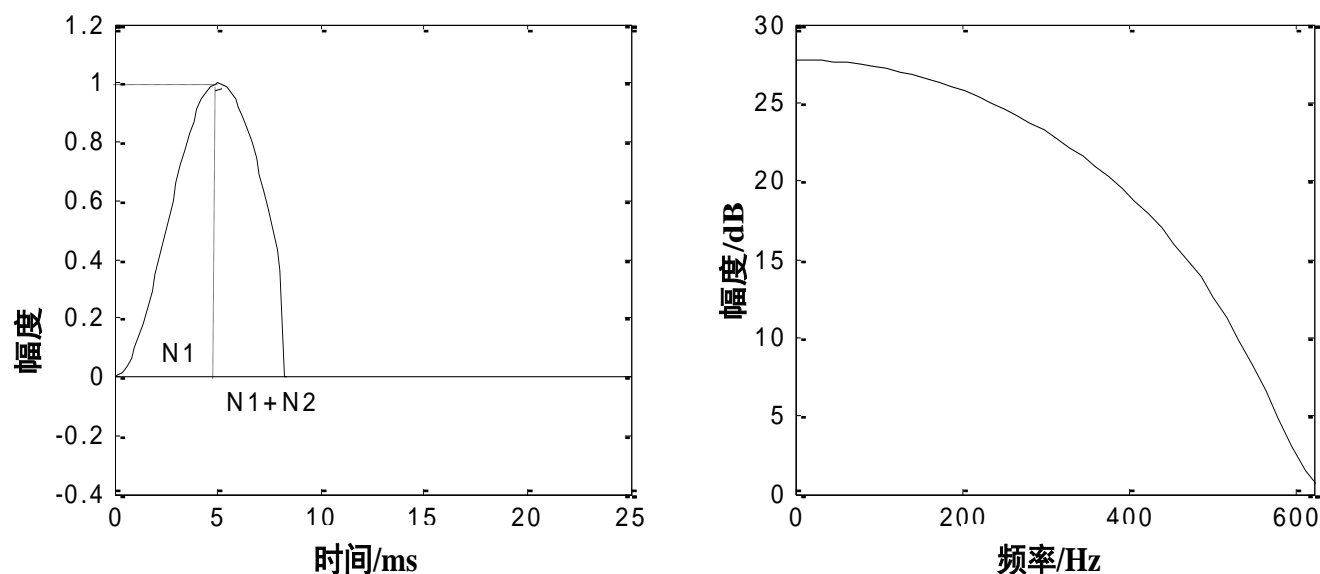


图2.9 三角波及其频谱图

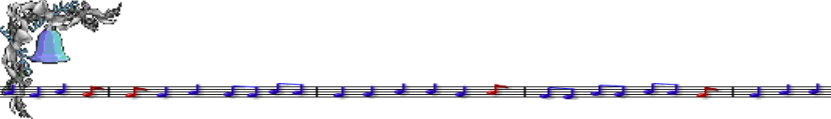


单个三角波的数学表达式为

$$g(n) = \begin{cases} \frac{1}{2} \left[1 - \cos \frac{n\pi}{N_1} \right] & 0 \leq n \leq N_1 \\ \cos \left[\frac{n - N_1}{2N_2} \pi \right] & N_1 \leq n \leq N_1 + N_2 \\ 0 & \text{其它} \end{cases}$$

其中： N_1 为斜三角波的上升时间
 N_2 为其下降时间





单个斜三角波的频谱 $G(e^{j\omega})$ 表现出一个低通滤波器的特性。其 z 变换的全极点形式为：

$$G(z) = \frac{1}{\left(1 - e^{-cT} \cdot z^{-1}\right)^2}$$

作为激励的斜三角波串可以用一串加了权的单位脉冲序列去激励单位斜三角波模型实现。这个单位脉冲串和幅值因子可以表示成下面的 z 变换形式

$$E(z) = \frac{A_v}{1 - z^{-1}}$$




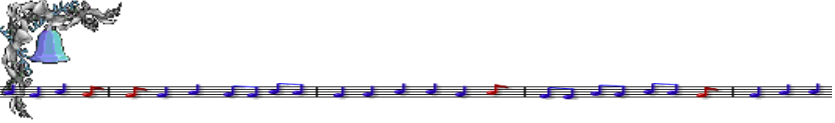


浊音激励模型可表示为

$$U(z) = E(z)G(z) = \frac{A_v}{1 - Z^{-1}} \cdot \frac{1}{\left(1 - e^{-cT} z^{-1}\right)^2}$$

清音可以模拟成随机白噪声。





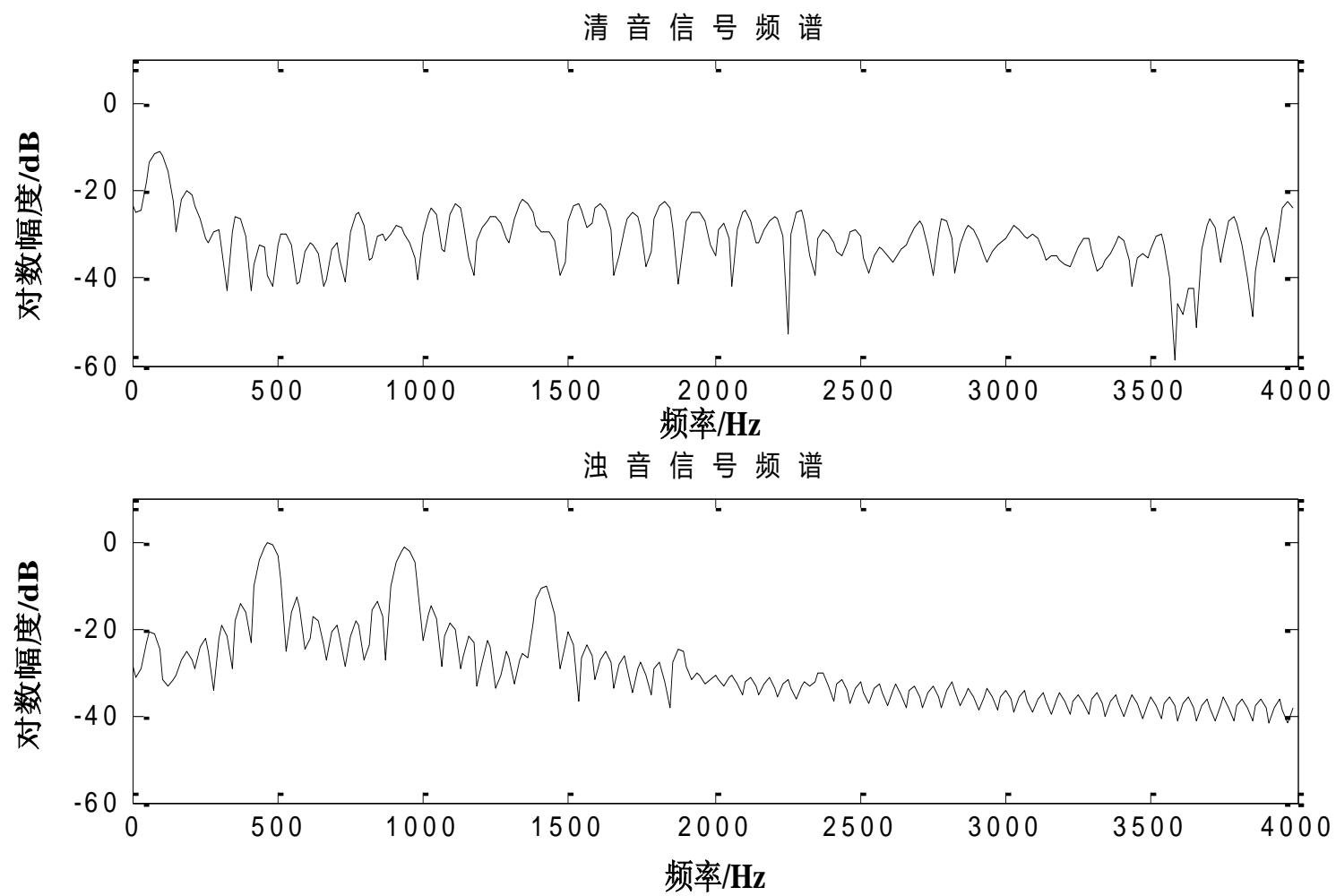
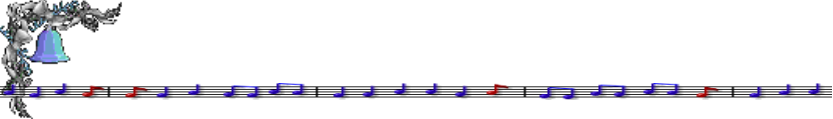
2.5.2 声道模型-（1）共振峰模型

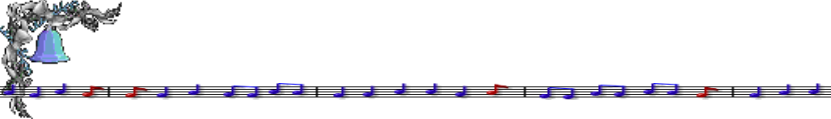
典型的声道模型有两种：无损声管模型和共振峰模型。

（1）共振峰模型

当声波通过声道时，受到声腔共振的影响，在某些频率附近形成谐振。反映在信号频谱图上，在谐振频率处其谱线包络产生峰值，一般把它叫作共振峰。







一个二阶谐振器的传输函数可以写成

$$V_i(z) = \frac{A_i}{1 - B_i z^{-1} - C_i z^{-2}}$$

实践表明，用前三个共振峰代表一个元音足够了。多个 V_i 叠加可以得到声道的共振峰模型：

$$V(z) = \sum_{i=1}^M V_i(z) = \sum_{i=1}^M \frac{A_i}{1 - B_i Z^{-1} - C_i Z^{-2}} = \frac{A}{1 - \sum_{k=1}^N a_k z^{-k}}$$





2.5.2 声道模型-（2）无损声管模型

无损声管模型：是假定声道由多个等长的不同截面积的管子串联而成的系统，并假定管子中的流体及管壁没有热传导和粘滞的损耗。在短时间内，声道可表为形状稳定的管道，并可以认为声波是沿管轴传播的平面波。



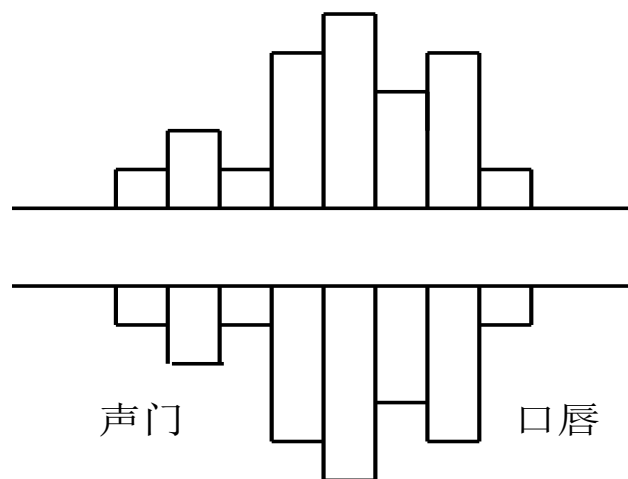
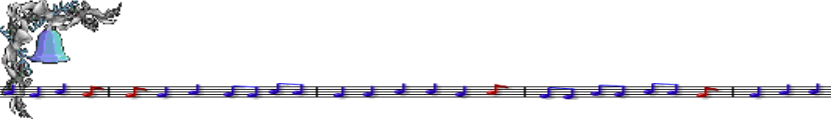


图 2.14 用声管模拟声道

对于N个无损声管级联的情况，可得到无损声管的传递函数为：

$$V(z) = \frac{G}{1 - \sum_{m=1}^N \alpha_m z^{-m}}$$





2.5.2 声道模型-（3）辐射模型

从声道模型输出的是速度波 $u_1(n)$ ，而语音信号是声压波 $P_1(n)$ 。二者倒比称为辐射阻抗 Z_1 ，它表征口唇的辐射效应。如果认为口唇张开的面积远远小于头部的表面积，利用单板开槽辐射的处理方法，可以得到辐射阻抗， r 近似为1

$$R(z) = R_0 (1 - rz^{-1})$$



由辐射引起的能量损耗正比于辐射阻抗的实部 $R(z)$ ，其频响曲线表现出一阶高通滤波器的特性。在实际信号分析时，常用所谓预加重技术。这样，模型只剩下声道部分，对参数分析就方便了。在语音合成时再进行解加重处理。





2.6 语音信号数字模型

2.6.1 数字模型

(1) 组成:

包括三部分：激励模型、声道模型和辐射模型。

激励源分浊音和清音两个分支，按照浊音/清音开关所处的位置来决定产生的语音是浊音还是清音。

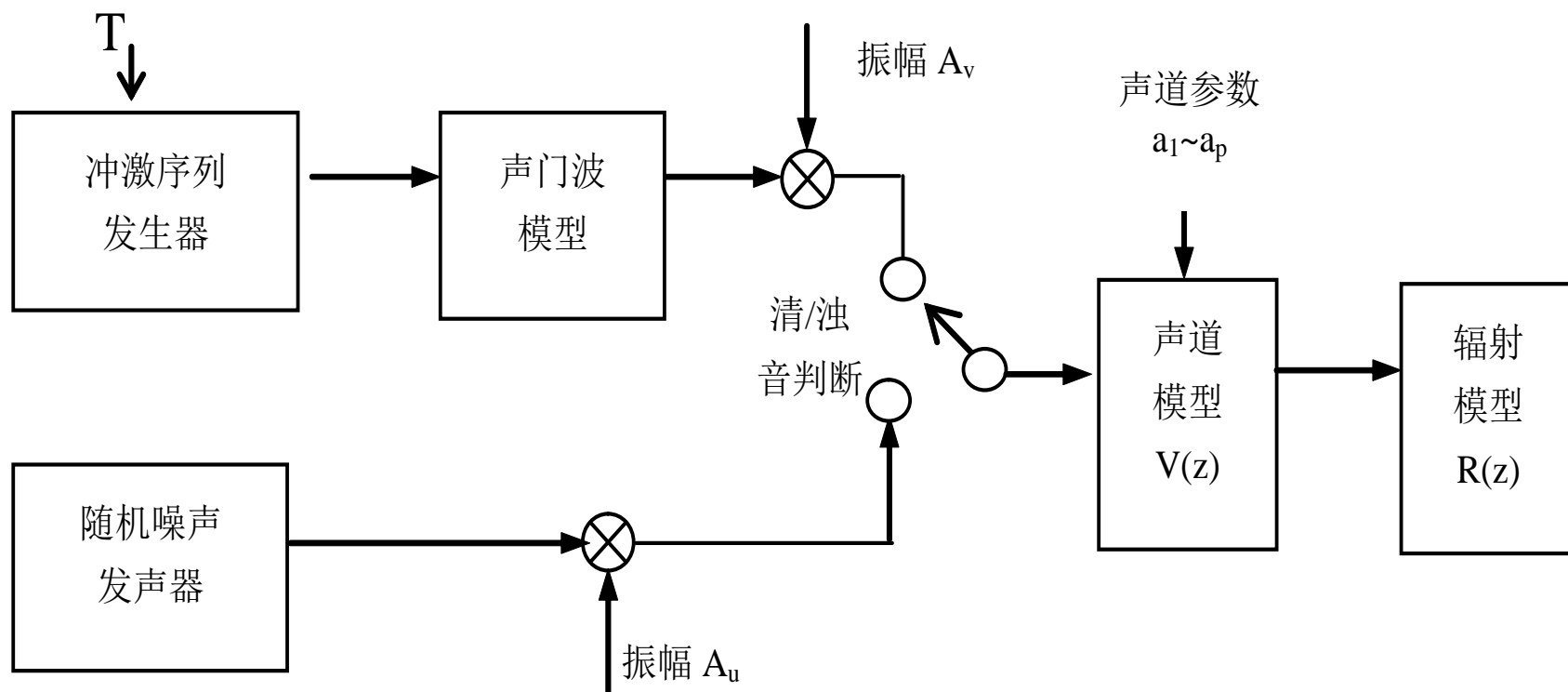


图 2.12 二元激励的语音生成模型





(2) 在浊音的情况下，激励信号由一个周期脉冲发生器产生。所产生的序列是一个周期为 T 的冲激序列, T 的倒数即为基音频率。为了使浊音的激励信号具有声门气流脉冲的实际波形，还需要使上述的冲激序列通过一个声门脉冲模型滤波器。





(3) 在清音的情况下，激励信号由一个随机噪声发生器产生。设其均值为0，方差为常数，幅度具有高斯概率分布。乘系数的作用是调节清音信号的幅度。

(4) 图2.16中画出了一段浊音语音产生过程中的有关波形。



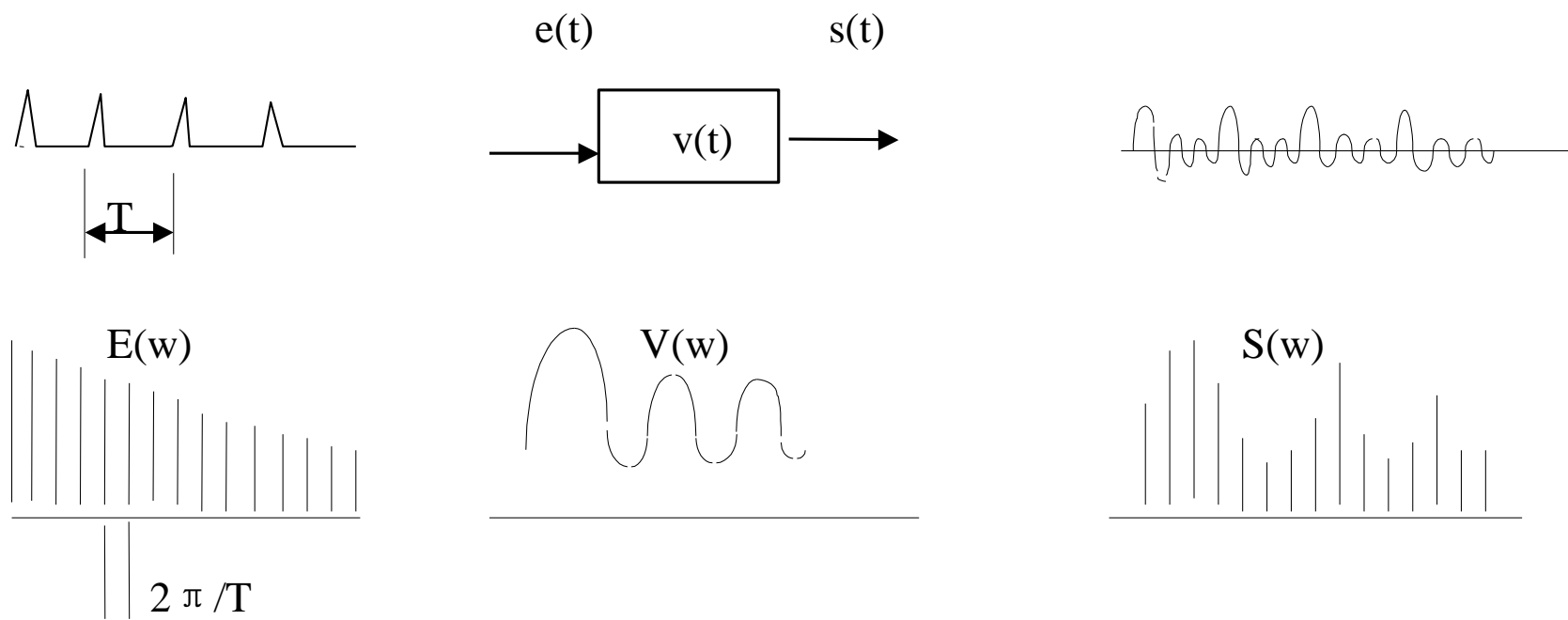
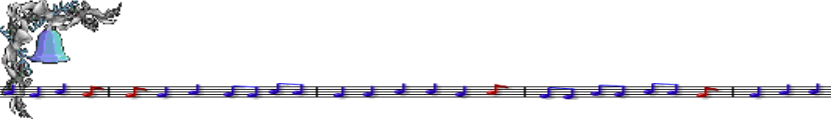


图 2.16 准周期脉冲序列激励声道产生浊音

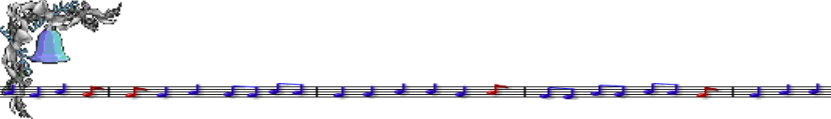




(5) 声道模型 $V(Z)$ 给出了离散时域的声道传输函数，把实际声道作为一个变截面声管加以研究，采用流体力学的方法可以导出，在大多数情况下它是一个全极点函数。 $V(Z)$ 可以表示为：

$$V(Z) = \frac{1}{\sum_{i=0}^P a_i Z^{-i}}, \quad a_0 = 1, a_i$$





把截面积连续变化的声管近似为P段短声管的串联，每段短声管的截面积是不变的。P称为这个全极点滤波器的阶。P值越大，模型的传输函数与声道实际传输函数的吻合程度越高。

辐射模型 $R(Z)$ 与嘴型有关，通常 $R(Z)$ 可以表示为

$$R(Z) = (1 - rZ^{-1}), r \approx 1$$





在这个模型中，除了 $G(Z)$ 和 $R(Z)$ 保持不变以外， T 、 A_v 、 A_u 、清/浊音开关的位置以及声道模型中参数 $a_1 \sim a_p$ 都是随时间而变化的，由于发音器官的惯性使这些参数的变化速度受到限制。对于声道参数，在 $10 \sim 30\text{ms}$ 的时间间隔内可以认为它们保持不变，因此，语音的短时分析是分帧进行的。对于激励源参数，多数情况下这一结果也是正确的。





(6) 统一的公式

离散时域语音信号 $s(n)$ 的 z 变换 $S(z)$ 可以用一个统一的公式来计算：

$$S(Z) = A * E(Z) * H(Z)$$

在浊音情况下， $E(z)$ 是一周期冲激序列的 z 变换且

$$A = A_v$$

$$H(Z) = G(Z)V(Z)R(Z)$$

则
$$S(Z) = A_v * E(Z) * H(Z)$$





在清音情况下， $E(z)$ 是一个随机噪声的 z 变换且

$$A = A_U \quad H(Z) = V(Z)R(Z)$$

则

$$S(Z) = A_u * E(Z) * H(Z)$$





2.6 语音信号数字模型- 模型局限性(1)

声道的传输函数具有全极点的性质，这对于元音和大多数辅音来说是比较符合实际的，但对于鼻音和阻塞音来说，由于出现了零点，这种模型就不够准确了。

一种解决问题的方案是在 $V(z)$ 中引入若干零点；另一种方法是适当提高阶数 P ，使得全极点模型能更好地逼近具有此种零点的传输函数。



2.6 语音信号数字模型- 模型局限性(2)

数字模型的基本思想是认为任何语音都是由一个适当的激励源作用于声道而产生的，这意味着激励源与声道系统是互相独立的。上述假定对于大多数语音是合适的，但在有些情况下，例如某些瞬变音，实际上声门和声道是互相耦合的，这便形成了这些语音的非线性特性。





2.6 语音信号数字模型- 模型局限性(3)

并非任何语音都能够明显地按清音和浊音来划分，有的音甚至也不是清音和浊音的简单叠加。
这种将语音信号截然分为周期脉冲激励和噪声激励两种情况的“二元激励”法在高质语音的合成中是不适用的。

