# 第一章　绪论

通过语音相互传递信息是人类最重要的基本功能之一。语言是人类特有的功能。声音是人类常用工具，是相互传递信息的最重要的手段。虽然，人可以通过多种手段获得外界信息，但最重要、最精细的信息源只有语言、图像和文字三种。与用声音传递信息相比，显然用视觉和文字相互传递信息，其效果要差得多。这是因为语音中除包含实际发音内容的语言信息外，还包括发音者是谁及喜怒哀乐等各种信息。所以，语言是人类最重要、最有效、最常用和最方便的交换信息的形式。另一方面，语言和语音与人的智力活动密切相关，与文化和社会的进步紧密相连，它具有最大的信息容量和最高的智能水平。

语音信号处理是研究用数字信号处理技术对语音信号进行处理的一门学科，处理的目的是用于得到某些参数以便高效传输或存储；或者是用于某种应用，如人工合成出语音、辨识出讲话者、识别出讲话内容、进行语音增强等。

语音信号处理是一门新兴的学科，同时又是综合性的多学科领域，是一门涉及面很广的交叉学科。虽然从事这一领域研究的人员主要来自信息处理及计算机等学科，但是它与语音学、语言学、声学、认知科学、生理学、心理学及数理统计等许多学科也有非常密切的联系。

语音信号处理是许多信息领域应用的核心技术之一，是目前发展最为迅速的信息科学研究领域中的一个。语音处理是目前极为活跃和热门的研究领域，其研究涉及一系列前沿科研课题，且处于迅速发展之中；其研究成果具有重要的学术及应用价值。

20 世纪 60年代中期形成的一系列数字信号处理方法和算法，如数字滤波器、快速傅里叶变换(FFT)等是语音信号数字处理的理论和技术基础。进入 70 年代之后，语音技术取得了许多实质性的进展：用于语音信号的信息压缩和特征提取的线性预测技术(LPC),已成为语音信号处理最强有力的工具，广泛应用于语音信号的分析、合成及各个应用领域；用于输入语音与参考样本之间时间匹配的动态规划方法。 80 年代初一种新的基于聚类分析的高效数据压缩技术——矢量量化(VQ)应用于语音信号处理中；而用隐式马尔可夫模型(HMM)描述语音信号过程的产生是 80年代语音信号处理技术的重大进展。近年来人工神经网络的研究取得了迅速发展，语音信号处理的各项课题是促使其发展的重要动力之一；同时，它的许多成果，也体现在有关语音的各项应用之中，尤其语音识别是神经网络的一个重要应用领域。

从技术角度讲，语音信号处理是信息高速公路、多媒体技术、办公自动化、现代通信及智能系统等新兴领域应用的核心技术之一。在高度发达的信息社会用数字化的方法进行语音的传送、存储、识别、合成、增强等是整个数字化通信网中最重要、最基本的组成部分之一。同时，由于语言是人类相互间进行沟通的最自然和最方便的形式，所以它是一种理想的人机通信方式，因而可为计算机、自动化系统等建立良好的人机交互环境，进一步推动计算机和其他智能机器的应用，提高社会信息化和自动化的程度。

语音处理技术的应用极其广泛，包括工业、军事、交通、医学、民用等各个领域。目前，语音处理技术处于蓬勃发展时期，已有大量产品投放市场，并且不断有新产品被开发研制，具有极其广阔的市场需要和应用前景。

在语音编码方面，如何在中低速率上获得高质量的语音，一直是其研究的主要目标。70年代中期，特别是80年代以来，语音编码技术有了突破性进展，提出了如多脉冲激励等一些有效的算法，产生了新一代的声码器，在16kbit/s以下速率上能够得到高质量的语音。

计算机和集成电路技术的发展，推动了语音信号处理的实用化。目前有很多专用语音处理芯片，这些芯片与微处理器或微型计算机相结合可以组成各种复杂的语音处理系统。其中语音合成在技术上比较成熟，在语音处理中影响也最大，上市产品多为有限词汇量的语音合成器。

然而，目前各种合成系统输出的语音音质跟自然语言的语音音质相差甚远，它并未真正解决机器说话的问题，因为其本质上只是一个不完满的声音还原过程。目前的合成只是停留在声道系统的发声过程上，其结果只是将书面语言转换成口头语言。而实现真正意义上的合成涉及到大脑的高级神经活动，而目前对这方面知道得还很少。同时，单音字节的声学语音学表现与音节在词语中有什么不同，尚未找到普遍的原则。由于涉及语言学、心理学和人脑的神经活动等问题，真正的语音合成问题尚处于研究阶段，这有待于信号处理、计算机方面和生理学、语言学、人工智能等领域的研究人员的共同努力。

在语音识别方面，很多专业人员对其理论和应用进行了广泛的研究，关于这方面的文献浩如烟海。目前，国内外有关论文每年达数千篇之多，但语音识别的研究比语音合成困难得多，其起步也较晚。

语音识别具有极广泛的应用领域，但它毕竟是一项综合性的、难度很大的高科技项目，从话语中提取满意的信息的过程是一项艰巨复杂的任务。虽然语音识别的研究已取得了很大进展，但还有很多困难甚至是原理性的问题有待解决。目前，语音识别领域的应用多是小词汇量特定人孤立词语音识别，是针对单讲话者的，能够得到较高的识别率。

在语音识别中，必然涉及到人是怎样从声音中提取信息和理解含义的问题。只有弄清人在收听声音时的生理过程并研究出模仿这些过程的模型，语音识别才可能得到一个飞跃的发展。如何充分借鉴和利用人在完成语音识别和理解时所利用的方法和原理就是一大课题，因而语音识别与人工智能之间有密切的联系。而目前只能从语音信号出发，用“隐过程”（如隐马尔可夫模型）来进行神经系统和听觉过程的模拟，是无法达到理想的识别和理解的效果的。

语音信号处理的理论和研究包括紧密结合的两个方面：一方面是从语音的产生和感知来对其进行研究，这一研究与语音学、语言学、认知科学、心理学和神经心理学等密不可分。另一方面是将语音作为一种信号来进行处理，包括传统的数字信号处理技术以及前面提到的一些新的应用于语音信号的处理方法及技术。

本书系统介绍了语音信号处理的原理、方法与应用，以及新方法和新技术。全书共分十五章，其中第二章介绍了语音处理需要的一些基础知识，包括语言和语音的基本特点；语音生成、语音感知等语音学、生理学和心理学基础。为了突出重点和节省篇幅，这一章只介绍与本书其余内容有直接关系的最基本的部分，如需进一步了解可参阅书中列出的参考文献。从第三章开始介绍语音的各种分析和处理技术，包括经典方法，如时域分析、频域分析等；以及各种新技术：同态处理、线性预测分析、矢量量化及隐马尔可夫模型技术等；还介绍了语音信号处理的各种应用，包括基音提取与共振峰估值、波形编码、声码器、语音合成、语音识别、说话人识别及语音增强等。

# 第二章　语音信号处理的基础知识

在研究分析各种语音信号处理技术及其应用之前，必须了解有关语音信号的一些基本特点；同时，要根据语音的产生过程建立一个既实用又便于分析的语音信号模型。这些都是语音信号处理的基础知识，对于语音信号处理的任何一个研究领域都是必需的，其中贯穿全书的是语音信号产生模型。

## 2.1　语音和语言

构成人类语音的是声音，然而这是一种特殊的声音，是由人讲话所发出的声音。语音是由一连串的音所组成，它是组成语言的声音。语音具有称为声学特征的物理物质。语音中各个音的排列由一些规则所控制，对这些规则及其含意的研究属于语言学的范酬，而对语音中音的分类和研究称为语音学。

人类生成语音过程的第一阶段是决定想传给对方的内容是什么，然后将内容转换成语言的形式。选择表现其内容的适当语句，将其按文法规则排列，便能构成语言的形式。由大脑对发音器官发出运动神经指令，发音器官各种肌肉运动，振动空气而形成语音波。这个过程可分为神经和肌肉的生理学阶段和产生语音波、传递语音波的物理阶段。

形成文章的基础是单词，单词简称词，是有意义的语言的最小单元。各单词由音节组成，音节又由音素组成。所谓音素是语言的元素即语言的最小基本单位，是发出各不相同音的最小单位；也就是说，音素都有其独立的各不相同的发音方法和发音部位，它是让听者能区别一个单词和另一个单同的声音的基础。音素分为两类：元音和辅音。在已知语言中元音有少至两个而多至 12 个，辅音从 10 多个至 70 多个。在英语中有 43 个音素。而音节的定义不一定明确，但是一个音节可以是 1 个元音和 1～2 个辅音组合。实际上，各种音素组合而构成语言时的连接方法有一几种限制，并不是所有的组合都存在。因此，一种语言中所用的音节数，远少于音素的组合数。

重音、语调和声调也是构成语言学的一部分，它们或者用来表示一句话中重要的单词，或者用来表示疑问句，或者用来表示说话人的感情。重音和语调是一种附加的信息，其中词的重音是西方语言如英语的一个重要特点，而语调实际上是讲话声音的调节，它决定于诸多因素，如语气、环境、讨论的话题等。语音中还有一个问题是同音异义词，它是指有相同的语音但是有两个或更多的不同意义。如汉语中的“语”与“雨”，英语中的“site”、“sight”、“cite”等就是同音异义词语音除了上述一些特点外，还存在所谓超语言学特点，如低语表示秘密、高声说话表示愤怒等。

对于我们所使用的汉语，有其特殊的、不同于英语的特点。汉语里也有元音和辅音的不同，其中不同的元音是由不同的口腔形状造成的，而不同的辅音是由发音部位和发音方法不同造成的。但是，汉语语音分析中总是把一个汉语音节分为声母和韵母两部分：声母就是一

个汉语音节开头的辅音，而韵母是汉字音节除了开头的声母以外的部分。在汉语中，有 21 个声母和 39 个韵母。

汉语的特点为汉语的自然单位是音节，每一个字都是单音节字，即汉语的一个音节就是汉语一个字的音，这里字是独立的发音单位。再由音节字构成词（其中主要是两音节字构成的词），最后再由词构成句子。而每一个音节字又都是由声母和韵母拼音而成；在音节中，声母比较简单，它们只是一个音素；而韵母则比较复杂。

汉语语音的另一个重要特点是它具有声调（即音调在发一个音节中的变化），这使它使用语声较其他语言更为经济。我国公布的汉语拼音方案中采用声调这个词。声调是一种音节在念法上的高低升降的变化。汉语有四种声调，即阴平（¯）、阳平（ˊ）、上声（ˇ）、去声（ˋ），上面括号内表示的是该声调的符号。由于有声调之分，所以参与拼音的韵母又有若干种（包括轻声在内至多有 5 种）声调。

汉语的特点是音素少、音节少。它大约有 64 个音素，但只有 400 个左右音节，即 400 个基本的发音。如考虑每个音节有 5 个声调，也只不过有 1 200 多个有调音节即不同的发音。

## 2.2 语音产生的过程及其声学特性

人的发音器官包括肺、气管、喉（包括声带）、咽、鼻和口等，如图 2-1 所示。这些器官共同形成一条形状复杂的管道，其中喉以上的部分称为声道，随着发出声音的不同其形状是变化的；而喉的部分称为声门。

产生语音的能量，来源于正常呼吸时肺部呼出的稳定气流，喉部的声带既是阀门，又是振动部件。在说话的时候，声门处气流冲击声带产生振动，然后通过声道响应变成语音。由于发不同的音时，声道的形状不同，所以听到不同的声音。以上就是发音器官发出声音时的大致情况。

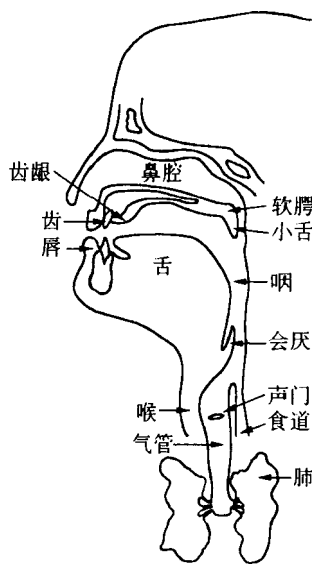


图 2-1 人的发音器官简图

喉部的声带是对发音影响很大的器官，声带的声学功能是为语音提供主要的激励源。由声带振动产生声音，是形成声音的基本声源。呼吸时左右两声带打开，讲话时则合拢起来。两声带之间的部位也称为声门。讲话时声带合拢因而受声门下气流的冲击而张开；但由于声带韧性迅速地闭合，随后又张开而闭合……声带开启和闭合使气流形成一系列脉冲。每开启和闭合一次的时间即振动周期称为音调周期或基音周期，其倒数称为基音频率，也简称为基频。声带振动的频率即基音决定了声音频率的高低，频率快则音调高，频率慢则音调低。基音的范围约为 70 ~ 350 Hz 左右，它随发音人的性别、年龄及具体情况而定，老年男性偏低，小孩和青年女性偏高。

语音由声带振动或不经声带振动来产生，其中由声带振动产生的音统称为浊音，而不由声带振动产生的音统称为清音。浊音中包括所有的元音和一些辅音，而清音中包括另一部分辅音。

声道是由咽、口腔和鼻腔组成，它是一根从声门延伸至口唇的非均匀截面的声管，其外形变化是时间的函数。成年男子声道的平均长度约 17 cm，而声道的截面积取决于其他发音器官的位置，它可以从零（完全闭合）变化到 $20cm^2$。声道是气流自声门声带之后最重要的、对发音起决定性作用的器官，发不同音时其形状变化是非常复杂的。

声道是一个分布参数系统，它有许多自然谐振频率（在这些频率上其传递函数具有极大值），所以声道是一谐振腔，它放大某一频率而衰减其他频率分量。谐振频率由每一瞬间的声道外形决定。讲话时，舌和唇连续运动，使声道常常改变外形和尺寸，随即改变谐振频率。如果声道的截面是均匀的，谐振频率将发生在

$$F_n = \frac{(2n-1)c}{4L} \qquad (n = 1,2,3\cdots)$$

式中，$c$ 为声速，在空气中为 $c = 350$ m/s；$L$ 为声管长度，$n$ 表示谐振频率的序号。如果 $L = 17$ cm，则谐振频率发生在 500 Hz 的奇数倍上，即 $F_1 = 500$ Hz，$F_2 = 1\,500$ Hz，$F_3 = 2\,500$ Hz，发元音 e[ə] 时声道截面最接近于均匀断面，所以谐振频率也最接近于上述值。而发其他音时，声道形状很少是均匀断面的，这些谐振点之间的间隔不同，但声道谐振点的平均密度仍然大约为每 1 kHz 有一个谐振点。

这些谐振频率称为共振峰频率，简称为共振峰，它是声道的重要声学特性。共振峰和声道的形状与大小有关，一种形状对应着一套共振峰。语音的频率特性主要是由共振峰决定的，当声音沿着声管传播时，其频谱形状就会随声管而改变。声门脉冲序列具有丰富的谐波成分，这些频率成分与声道的共振频率之间相互作用的结果对语音的音质有很大影响。由于声道的大小随不同讲话而不同，因此共振峰频率与讲话者有密切关系。即使是音素相同，但因讲话者不同，共振峰也有相当大的变化。

共振峰用依次增加的多个频率表示，如 $F_1$、$F_2$ 等，称为第一共振峰、第二共振峰等。在声学语音学中通常考虑 $F_1$ 和 $F_2$，但在语音识别技术中至少要考虑三个共振峰，而在语音合成技术中考虑五个共振峰是最为现实的。表 2-1 给出了前三个共振峰的大致范围（单位为 Hz），这些数值只是概略的，因为不同的人特性变化相当大。

表 2-1　前三个共振峰的频率范围

| | 频率范围 /Hz | | |
|---|---|---|---|
| | 成年男子 | 成年女子 | 带　宽 |
| $F_1$ | 200 ~ 800 | 250 ~ 1 000 | 40 ~ 70 |
| $F_2$ | 600 ~ 2 800 | 700 ~ 3 300 | 50 ~ 90 |
| $F_3$ | 1 300 ~ 3 400 | 1 500 ~ 4 000 | 60 ~ 180 |

根据上面的介绍，语音的特性完全由声门、声道和口鼻决定。

## 2.3　语音感知

发音的目的就是让对方听懂并且理解这种声音，因而发音的方法与人的听觉能力密切相关。人的听觉能力，既有人工智能无法模仿、高能力的一方面，也有无能为力的一方面。所谓高能力是指，即使众多讲话者以各种声音、习惯、方言、语调同时讲话，甚至其中一些人讲话含糊不清，听话者也能准确地听懂应该听取的声音；所谓无能为力的一面是指，频率相近

或间隔太短的声音无法区别。另外，也存在这种现象，即两个音同时存在时，一方的音被另一方的音掩盖。所以听觉能力中包含着大脑的高度发达的语言理解能力。

由发音器官产生的声波传到收听者的耳中，收听者的听觉器官运动，作为神经脉冲经听觉神经传播到大脑中。这样，说话方想表达的语言信息被对方理解。语音波不仅传到对方的耳中，同时也传到说话人本人的耳中。说话人边听到自身反馈的声音，边不断地对发音器官进行调节。以上所说的语音的形成和接收紧密相连，称之为“语言通道”，如图 2-2 所示，它是由语言学、生理学及物理学三个阶段组成。

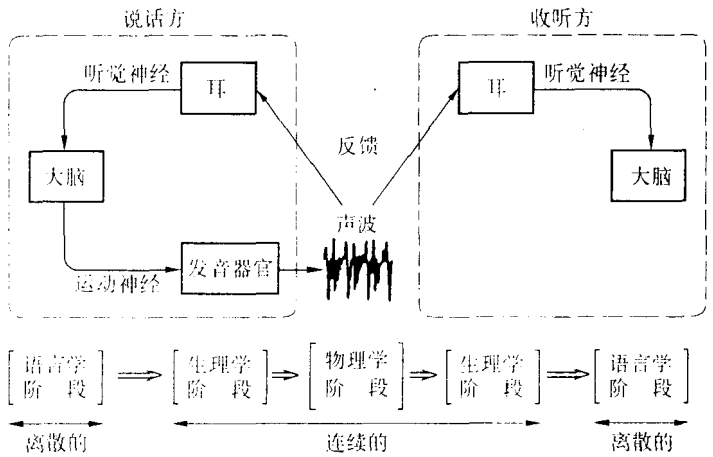



图 2-2　语言通道

## 2.4　语音信号的数字模型

这里，表示抽样语音信号的离散模型是特别重要的。为了定量描述语音处理所涉及到的某些因素，虽然已经假定了许多不同的模型，但是可以肯定，目前还没有发现一种可以详细描述人类语音中已观察到的全部特征的模型（由于它的复杂性，也许不可能找到一个理想的模型）建立模型的基本准则是要寻求一种可以表达一定物理状态下的数学关系，要使这种关系不仅具有最大的精确度，而且还要最简单。

我们希望模型既是线性的又是时不变的，这是最理想的模型。但是语音信号是一连串的时变过程，根据语音的产生机遇，不能精确地满足这两种性质此外，声门和声道相互耦合，还形成语音信号的非线性特性。然而，作出一些合理的假设，在较短的时间间隔内表示语音信号时，可以采用线性时不变模型。下面将给出经典的语音信号数字模型，这里，语音信号被看成是线性时不变系统（声道）在随机噪声或准周期脉冲序列激励下的输出。这一模型用数字滤波器原理加以公式化后，就成为本书其余部分讨论语音处理技术的基础。

长期研究证实，发不同性质的音时，激励的情况是不同的，大致分为两大类：① 发浊音时。此时气流在通过绷紧的声带时，冲激声带产生振动，使声门处形成准周期性的脉冲串，并用它去激励声道。声带绷紧的程度不同时，振动频率也不同。该频率就是音调频率，其倒数为音调周期。不同人的音调周期是不同的，男子大，女子小；老人大 ,小孩低。② 发清音时。此时声带松弛而不振动，气流通过声门直接进入声道。

产生语音信号的框图如 2-3所示。

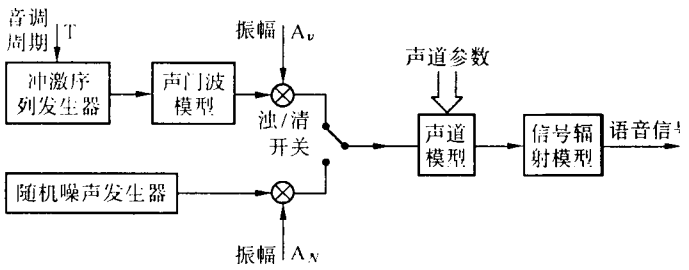



图 2-3　语音信号的产生模型

下面讨论语音信号的数字模型。

### 2.4.1　激励模型

发浊音时，由于声带不断张开和关闭，将产生间歇的脉冲波。根据测量结果，这个脉冲波类似于斜三角形的脉冲，如图 2-4(a) 所示。因此，此时的激励信号是一个以基音周期为周期的斜三角脉冲串。单个三角波形的数学表达式如下

$$g(n)=\begin{cases}\dfrac{1}{2}\left[1-\cos\left(\dfrac{\pi n}{N_1}\right)\right] & (0\leqslant n\leqslant N_1)\\ \cos\left[\pi\left(\dfrac{n-N_1}{2N_2}\right)\right] & (N_1\leqslant n\leqslant N_1+N_2)\\ 0 & \text{其他}\end{cases}\tag{2-1}$$

式中，$N_1$ 为斜三角波上升部分的时间，$N_2$ 为其下降部分的时间

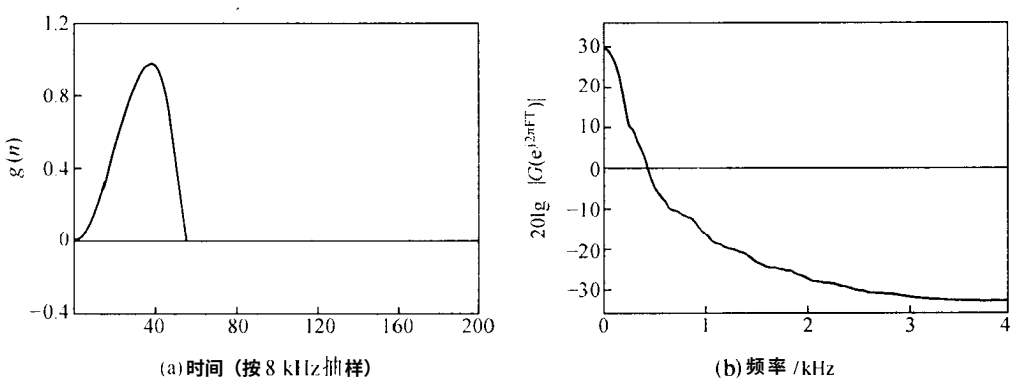



图 2-4　单个斜三角波及其频谱

单个斜三角波形的频谱 $G(e^{j\omega})$ 如图 2-4(b) 所示。由图可见，它是一个低通滤波器。通常更希望将其表示为 Z 变换的全极模型的形式

$$G(z)=\frac{1}{(1-e^{-CT}z^{-1})^2}\tag{2-2}$$

这里，$C$ 是一个常数。显然，上式表明斜三角波可描述为一个二极点的模型。

因此，斜三角波脉冲可看作加权的单位脉冲串激励上述单个斜三角脉冲的结果。而该单

位脉冲串及幅值因子可表示成下面的 Z 变换形式

$$E(z) = \frac{A_v}{1 - z^{-1}} \tag{2-3}$$

所以整个激励模型可表示为

$$U(z) = G(z)E(z) = \frac{A_v}{1 - z^{-1}} \cdot \frac{1}{(1 - e^{-CT}z^{-1})^2} \tag{2-4}$$

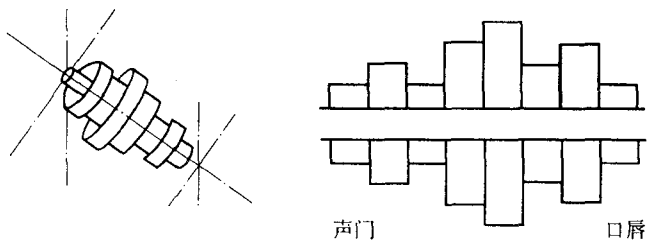
另一种是发清音的情况。这时声道被阻碍形成湍流，所以可模拟成随机白噪声。实际上可使用均值为 0、方差为 1，并在时间或在幅度上为白色分布的序列。

应该指出，这样简单地把激励分为浊音和清音两种情况是不严格的。对于某些音，即使是把两种激励简单地叠加起来也是不合适的。但是，若将这两种激励源经过适当的网络后，是可以得到良好的激励信号的。为了更好地模拟激励信号，有人提出在一个音调周期时间内用多个斜三角波（例如三个）脉冲的方法；此外，还有用多脉冲序列和随机噪声序列的自适应激励的方法等。

关于声道部分的数学模型，目前有两种观点：一是将声道视为由多个不同截面积的管子串联而成的系统，由此推导出“声管模型”；一是将声道视为一个谐振腔，由此推导出“共振峰模型”。下面分别介绍。

### 2.4.2 声管模型

最简单的声道模型是将其视为由多个不同截面积的管子串联而成的系统，这就是声管模型。在语音信号的某一“短时”期间，声道可表示为形状稳定的管道，如图 2-5 所示。




(a) 立体图　　(b) 断面图

图 2-5　声道的声管模型

在声管模型中，每个管子可看作为一个四端网络，这个网络具有反射系数，这些系数和第六章中将要介绍的线性预测的参数之间有唯一的对应关系。这时声道可由一组截面积或一组反射系数来表示。

通常用 $A$ 表示声管的截面积。由于语音的短时平稳性，假设在短时间内，各段管子的截面积 $A$ 是常数。设第 $m$ 段和第 $m+1$ 段的声管的截面积分别为 $A_m$、$A_{m+1}$，设 $k_m = (A_{m+1} - A_m)/(A_{m+1} + A_m)$，称为“面积和差比”，其取值范围为 $-1 < k_m < 1$。它实际上是线性预测的反射系数。

### 2.4.3 共振峰模型

另一种声道模型是将其视为一个谐振腔，共振峰就是这个腔体的谐振频率。由于人耳听

觉的纤毛细胞就是按频率感受而排列其位置的，所以这种共振峰的声道模型方法是非常有效的。实践表明，用前三个共振峰来代表一个元音就足够了；对于较复杂的辅音或鼻音，大概要用到五个以上的共振峰才行。

基于共振峰理论，可以建立三种实用的模型：级联型、并联型和混合型。

1. 级联型

此时认为声道是一组串联的二阶谐振器。根据共振峰理论，整个声道具有多个谐振频率和多个反谐振频率，所以它可被模拟为一个零极点的数学模型；但对于一般元音，可以用全极点模型。其传输函数为

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2-5}$$

式中，$N$是极点个数，$G$是幅值因子，$a_k$是常系数。此时可将此传输函数分解为多个二阶极点的网络的串联，则得

$$V(z) = \prod_{k=1}^{M} \frac{1 - 2e^{-B_k T}\cos(2\pi F_k T) + e^{-2B_k T}}{1 - 2e^{-B_k T}\cos(2\pi F_k T) z^{-1} + e^{-2B_k T} z^{-2}} = \prod_{k=1}^{M} \frac{a_k}{1 - b_k z^{-1} - c_k z^{-2}} \tag{2-6}$$

式中

$$\begin{aligned} c_k &= \exp(-2B_k T) \\ b_k &= 2\exp(-\pi B_k T)\cos(2\pi F_k T) \\ a_k &= 1 - b_k - c_k, a_1 a_2 \cdots a_M = G \end{aligned} \tag{2-7}$$

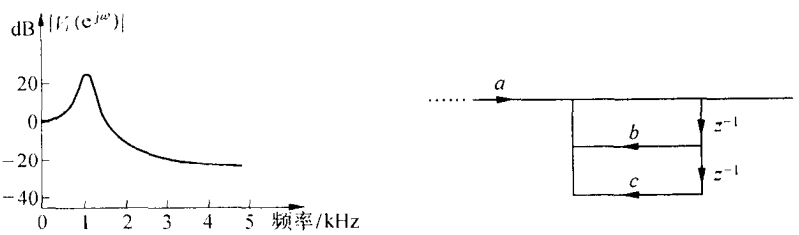
$M$为小于$(N+1)/2$的整数。

若$z_k$是第$k$个极点，则有$z_k = e^{-B_k T} e^{-2\pi F_k T}$，其中$T$为取样周期。

取式(2-6)中的某一级，设为

$$V_k(z) = \frac{a_k}{1 - b_k z^{-1} - c_k z^{-2}} \tag{2-8}$$

则可画出系统流图和幅频特性，如图2-6所示。




(a) 二阶谐振器的幅频特性　　(b)二阶谐振器的流图

图 2-6　二阶谐振器

而整个级联型模型框图可表示为图2-7的形式。如$N = 10$，则$M = 5$。此时整个声道可模拟成图2-8的形式，图中$G$是幅值因子。

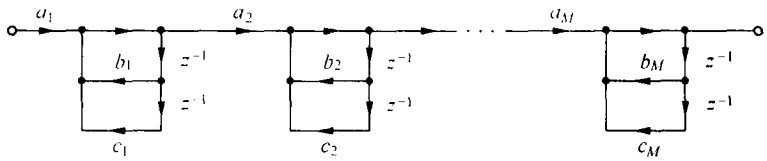



图 2-7　级联型系统框图

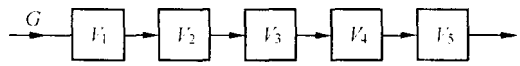



图 2-8　级联型共振峰模型

2. 并联型

对于非一般的元音和大部分辅音，必须采用零极点模型。此时其传输函数为

$$V(z) = \frac{\sum_{r=0}^{R} b_r z^{-r}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2-9}$$

通常，$N > R$，且设分子与分母无公因子及分母无重根，则上式可分解为部分分式之和

$$V(z) = \sum_{k=0}^{R} \frac{A_k}{1 - B_k z^{-1} - c_k z^{-2}} \tag{2-10}$$

这就是并联型的共振峰模型，如图 2-9 所示（$M = 5$）。

3. 混合型

上面两种模型中，级联型比较简单，可用于描述一般的元音。级联的级数取决于声道的长度。当声道长度为 17 cm左右时，取 3 ~ 5 级即可。当鼻化元音或鼻腔参与共振，以及发阻塞音或摩擦音等时，用级联型描述就不合适了；此时腔体具有反谐振特性，必须考虑加入零点，使之成为极零点模型。为此可采用并联型结构。它比级联型复杂些，每个谐振器的幅度都要独立控制。

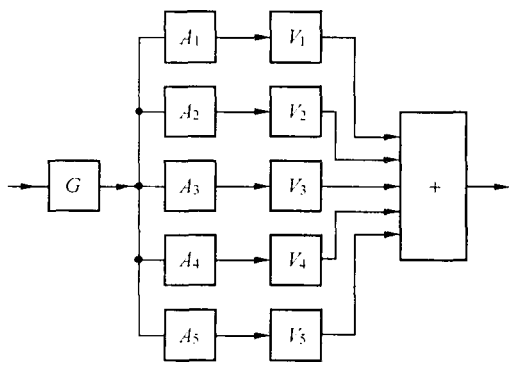



图 2-9　并联型共振峰模型

将级联型和并联型结合起来的混合型也许是比较完备的一种共振峰模型，如图 2-10 所示。该模型能够根据不同性质的语音进行切换。图中的并联部分，从第一到第五共振峰的幅度都可以独立地进行控制和调节，用来模拟辅音频谱特性中的能量集中区。此外，并联部分还有一条直通路径，其幅度控制因子为 $AB$，这是专为一些频谱特性比较平坦的音素（如[f]、[p]、[b] 等）而考虑的。

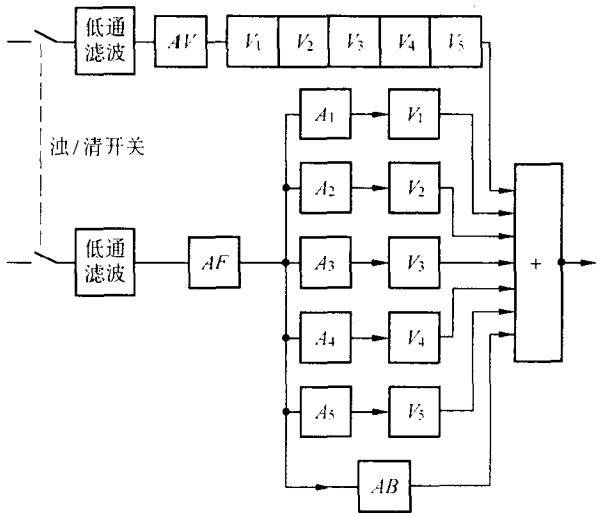



图 2-10　混合型共振峰模型

### 2.4.4　辐射模型

声道的终端为口和唇。从声道输出的是速度波，而语音信号是声压波，二者之倒比称为辐射阻抗 $Z_L$。它表征口和唇的辐射效应，也包括圆形的头部的绕射效应等。

研究表明，口唇端辐射在高频端较为显著，在低频端时影响较小，所以辐射模型 $R(z)$ 应是一阶类高通滤波器的形式。口唇的辐射效应可表示为

$$R(z) = R_0(1 - z^{-1}) \tag{2-11}$$

它是一阶后向差分。

在语音信号模型中，如果不考虑冲激脉冲串模型 $E(z)$，则斜三角波模型是二阶低通，而辐射模型是一阶高通，所以实际信号分析中常采用“预加重技术”。即在对信号取样之后，插入一个一阶的高通滤波器，这样，只剩下声道部分，就便于对声道参数进行分析了。在语音合成时再进行“去加重”处理，就可以恢复原来的语音。常用的预加重因子为 $1 - [R(1)/R(0)] \cdot z^{-1}$。这里，$R(n)$ 是语音信号的自相关函数。通常对于浊音，$[R(1)/R(0)] \approx 1$；而对于清音，该值可取得很小。

由上面所述，完整的语音信号数字模型可以用三个子模型：激励模型、声道模型和辐射模型的串联来表示。其转移函数为

$$H(z) = U(z)V(z)R(z) \tag{2-12}$$

这里，$U(z)$ 是激励信号 —— 声门脉冲即斜三角波的形式，$V(z)$ 是声道传递函数，既可以用声管模型，也可以用共振峰模型来描述。在共振峰模型中，又可采用级联型、并联型或混合型等几种形式。

另外，语音信号产生模型可以用图 2-3 表示。发不同性质的音时，激励的情况是不同的。利用浊音和清音激励发生器二者的切换，就可以模拟激励形式的改变。

应该指出，式(2-12)所示模型的内部结构并不和语音产生的物理过程相一致，但这种模型和真实模型在输出处是等效的。另外，这种模型是“短时”的模型，因为一些语音信号的

变化是缓慢的，例如元音在 10 ~ 20 ms 内其参数可假定不变。这里声道转移函数 $V(z)$ 是一个参数随时间缓慢变化的模型。另外，这一模型认为语音是声门激励线性系统 —— 声道所产生的；实际上，声带 - 声道互相作用的非线性特性还有待研究。同时，正如 2.4.1 中所指出的，模型中用浊音和清音这种简单的划分方法是有缺陷的，对于某些音是不适用的，例如浊音当中的摩擦音。这种音要有发浊音和发清音的两种激励，而且二者不是简单的叠加关系。对于这些音可用一些修正模型或更精确的模型来模拟。

## 2.5 语音信号的统计特性

对语音信号统计特性的研究表明，其幅度分布的概率密度有两种近似表达式。较好的是修正伽玛(Gamma) 概率密度

$$f(x) = \frac{\sqrt{k}}{2\sqrt{\pi}} \cdot \frac{e^{-k|x|}}{\sqrt{|x|}} \qquad (2\text{-}13)$$

精度稍差一点的是拉普拉斯(Laplacian) 分布

$$f(x) = 0.5\alpha e^{-\alpha|x|} \qquad (2\text{-}14)$$

这些密度曲线示于图 2-11 中，图中还给出了一段天气预报语音的幅度直方图（该话音内容为："Clear and cold tonight, low in the upper teens in the city, ten above in the suburbs")。图中还画出了高斯密度曲线以便比较。注意语音主要集中在幅度较小的区域。

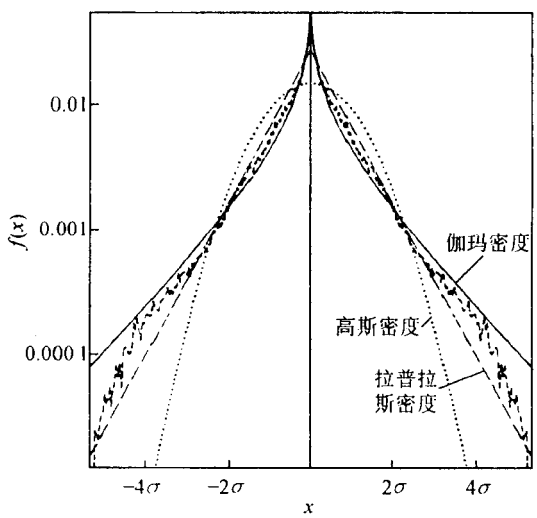



图 2-11 修正伽玛密度（实线）、拉普拉斯密度（虚线）和高斯密度（点线）以及天气预报语音的长期幅度分布（不规则的虚线）

图 2-12 表示出英语和日语发音十多分钟所得到的声音振幅累计分布。图中，横轴所表示的是以长时间有效值为相对基准的振幅。因为从振幅大的点累积，所以，纵轴表示超过这个振幅值的频度。由图可见，不论日语还是美国英语，其动态范围都超过 50 dB。

## 2.6 语音特性分析实例

下面以元音为例讨论一下其波形的性质。这些性质在后面要经常地引用。

因为元音属于浊音，所以其声门波形为图 2-13 所示的脉冲序列，脉冲之间的间隔为基音周期，这个函数用 $g(t)$ 表示。将它加于声道，得到的语音信号是 $g(t)$ 与声道冲激响应 $h(t)$ 的卷积。这里假定 $g(t)$ 不受声道形状影响。假定声道传递函数是全极点的，其脉冲响应就是一系列衰减的正弦波之和，$H(z)$ 的每一个极点对应一个衰减振荡，得到的典型时间函数如图 2-14 所示。每个高峰代表一个新的声门脉冲的起点，因此，它们之间的间隔等于声门脉冲的周期。

下面考察其频域特性。脉冲序列具有丰富的谐波。如果将 $g(t)$ 考虑为脉冲序列与声门

脉冲波形的卷积，得到的频谱就是间隔等于基音频率的脉冲序列与声门波形的傅里叶变换的乘积。这种变换通常有复数零点落于我们关心的频率范围内，这些零点在与共振峰相互作用的激励频谱包络中产生最小点。随着基音、说话条件、说话人不同及其他条件的变化，声门脉冲形状有很大的变化，因此准确的零点位置难以确定。通常大约在 0.8 ~ 1.0 kHz 以上用 12 dB/ 倍频程的下降来表示这种影响。

还有一个因素要考虑。由于话音以嘴唇辐射出去时其声压与口腔中体速度的微分成正比，这使话音频谱的幅度有 6 dB/ 倍频程的提升。通常，把这种提升的影响与声门的影响结合

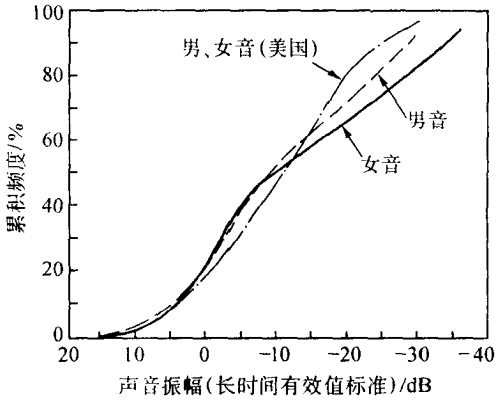



图 2-12　语音振幅的累计频度分布

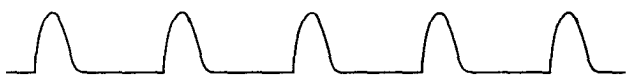


图 2-13　声门脉冲序列

图 2-14　声道对声门脉冲响应的输出

起来，以便于研究声道滤波器，采用 6 dB/ 倍频程滚降的脉冲序列作为“虚拟”的激励频谱。由于加窗是实际谱分析的一部分，所以图 2-15 表示的是加窗之后的激励频谱。

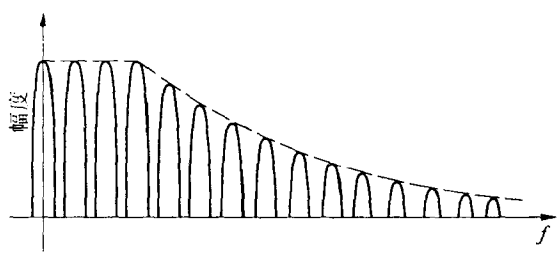



图 2-15　理想的声门脉冲序列频谱

将这样得到的激励加于声道。设上述所示的有效频谱为 $G(f)$，声道传递函数为 $H(f)$，则输出频谱为 $G(f)H(f)$。$H(f)$ 的特点是最大值与共振峰相对应，见图 2-16。

输出语音频谱如图 2-17 所示，其中虚线称为谱包络，其形状是由 $H(f)$ 和 $G(f)$ 的包络乘积得到的。恢复这个谱包络是许多语音处理应用中的主要问题，因为正是谱包络携带了主要的发音信息。第六章介绍的线性预测技术之所以非常重要，正是由于它所提供的谱包络分析方法是快速、准确，并且在理论上完全得到证明的方法。

为了直观地了解语音的特性，图 2-18 给出了一段语音的时间波形。它是摘自天气预报

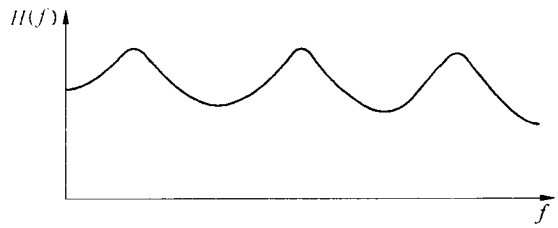



图 2-16　声道频率响应，最大值与共振峰相对应

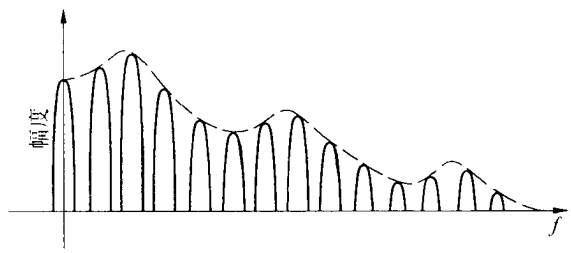



图 2-17　输出语音频谱

中的一个句子，该句话为“ten above in suburbs”。对这段语音以每秒 8 000个样点进行采样；时间延迟在图中用相等间隔来表示。

由图可见，在不同的音素之间实际上没有明显的分界，几乎每个音素都逐渐消失在其后面的音素中。图中对音素的描述都是从一定时刻开始并用大写英文字母表示，但这只是大致的位置。

[t]音的开始大约发生在 7 s这一时刻，图中用 A 点表示。由 B点开始是“ten ”中的[ɛ]音。这时可看到语音波形特有的形式：每个周期开始都有一个明显的高峰，接着是一串衰减振荡。开始的高峰是由声门脉冲的起点造成的，接着的激励振荡是口腔谐振系统的脉冲响应。在时标 7.10和 7.15 之间大约有 7.5个周期；因此说话人的基音大约为 150 Hz ,这个数值对男性发音来说是合理的。

[n]音由 C 点开始延迟约 4 个周期到 D 点。紧接着是“above’中的 [ə]音，在这两个单词之间没有将它们分开。[ə]音长度约为 5 个或 6个周期，[b]音大约在 E点开始。振荡一直持续到[b] 音发出。后面的[ʌ] 音在 F 点开始，一直持续到 G 点。

图中词组“in the suburbs ”由 H点开始。在由“in ’中的[n] 音向 “the ”中 的 [ð]音转音的过程中，可以看到协同发音的例子。在口张开发“the ’中 的 “ð” 音之前，[n] 音的波形实际上是保持不变的；[ð]音开始只有一种低电平噪声加于 [n]音的最后两三个周期上，并使[ə] 音的头一两个周期的波形稍微有些起伏。这里的“th” 音实际上几乎完全简化为一个[ə] 音，英语中的大多数“the ’都是这样。“suburbs” 中的前一个[s] 音从 K 点一直持续到 L 点。

Q 点以后完全无声使句子没有明显的终止点。但有时情况也并不如此，因为人们说话时经常都控制自己的呼吸，一直到一句话说完才透过一口气，从而给语音识别中的端点检测造成困难。此外在图中还可看出，8.35 秒以后说话人为准备下一句话呼气而产生的逐渐增大的噪声波形。

图 2-19 给出 “above” 中[ʌ] 音的傅里叶变换，时间大约在图 2-18 中 7.45 处开始。取时间波形宽度为 256 个样本，大约包括 4 个基音周期。在频谱图中基音的谐波表示得很清楚。

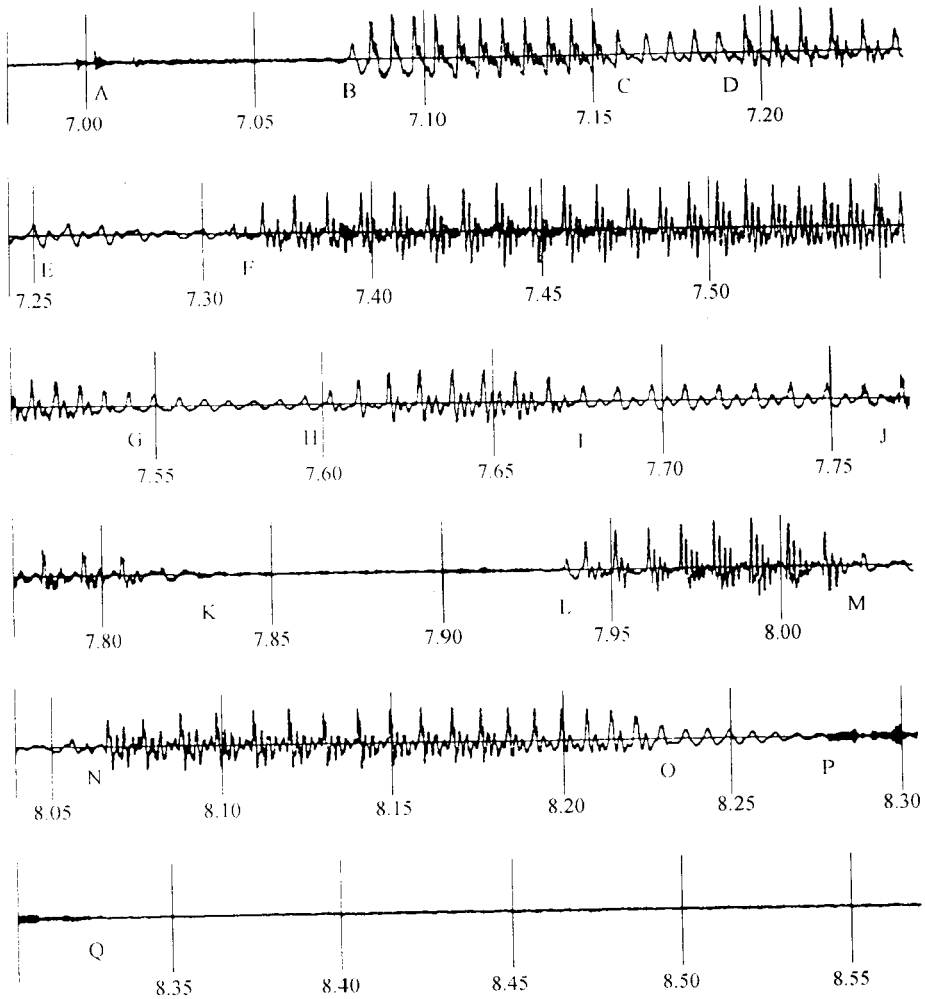



图 2-18　一段语音的时域波形

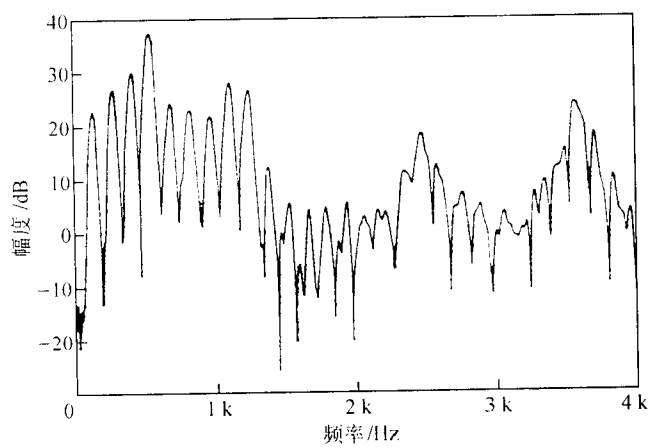



图 2-19　元音 [ʌ] 的频谱

在 0 ~ 1 500 Hz之间几乎有 11 个峰点，因此基音约为 136 Hz。观察图 2-18 中周期之间的距离可以证明这里的推算是正确的。在图 2-18 中，在 7.45 和 7.50之间约 6.5个周期，由此可估计基音约为 130 Hz。这两种结果是相当一致的。频谱表示能量在 550、1 150、2 450、3 600 Hz 各频率附近最为集中，这些频率就是共振峰。前三个共振峰的频率数值也与表 2-1 中 [ʌ] 音的共振峰频率数值一致。

单词“suburbs”中开始的 [s] 音的傅里叶变换示于图 2-20 中。可以看出 [s] 音中不乏高频能量。图中可见频谱峰点之间的间隔是随机的，表明 [s] 音中没有周期分量，这与原来的预计是一样的。

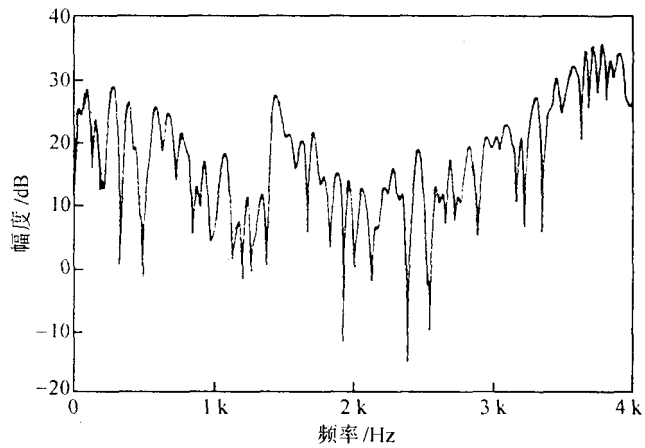



图 2-20 辅音[s] 的频谱

## 2.7 语音的质量评价

用语言进行信息交换，必须以听到的语音质量作为评价标准。实际上，音质都是采用主观评价方法。

对语音通信系统的总评价是对所收到的语音信号的质量评价，即所传出的语音能正确无误接收的比例，称这个量为语音清晰度或可懂度。理论上，语音质量测试可分为两类：一类是音节以下（如音素、声母、韵母）的语音单元的测试，这常称为“清晰度”测试；清晰度测试可以元音、辅音为基础或以声母、韵母为基础，再根据音节成分算出音节清晰度。另一类是音节以上（如词、句）的语音单元的测试，常称为“可懂度”测试。这些测试的基本原理是相同的，只是测试的单元有所不同；而且可以按条件用公式从小单元的清晰度计算出大单元的可懂度来。这些测试都属于主观听觉的测试方法，即以人的测听效果为依据。

在测试汉语时，可采用 1984 年颁布并实行的中华人民共和国电子工业部部颁标准 SJ 2467—84：通信设备汉语清晰度测试方法。它是汉语清晰度的测试标准，是专为评定语言通信设备和其他语言传输系统的话音清晰度而制定的。

在美国也有类似的测试标准： DRT(Diagnostic Rhyme Test，押韵测试）和 DAM(Diagnostic

Acceptability Measure Test,可接受程度测试)。其中 DRT 用于可懂度测试，它采用数对单词每对中的单词仅字首辅音不同，而每个辅音仅有一个参数不同。

## 参 考 文 献


1 G.Fant.Acoustic Theory of Speech Production. Mouton,The Hague,1970

2 J.L.Flanagan.Speech Analysis,Synthesis and Perception. New York:Springer—Verlag,2nd Ed.,1972

3 L.R.Rabiner and R.W.Schafer.Digital Processing of Speech Signals. Prentice Hall Inc., 1978

4 马大猷.语言信息和语言通信.上海：知识出版社,1987

5 中华人民共和国电子工业部部颁标准 SJ 2467—84.通信设备汉语清晰度测试方法.1984